NASA CONTRACTOR REPORT 177388

JAN 1986
RECEIVED
NASA STI FACILITY
ACCESS DEPT.

Effect of Time Span and Task Load on Pilot Mental Workload

S. L. Berg
T. B. Sheridan

(NASA-CR-177388) EFFECT OF TIME SPAN AND TASK LOAD ON PILOT MENTAL WORKLOAD (Massachusetts Inst. of Tech.) 152 p
HC A08/MF A01 CSCL 05H
N86-16901
Unclas
G3/54 03695

CONTRACT NAG2- 227
January 1986

NASA

NASA CONTRACTOR REPORT 177388

Effect of Time Span and Task Load on Pilot Mental Workload

S. L. Berg
T. B. Sheridan

Massachusetts Institute of Technology
Cambridge, Massachusetts

Prepared for
Ames Research Center
under Grant NAG 2-227

NASA
National Aeronautics and
Space Administration

Ames Research Center
Moffett Field, California 94035

EFFECTS OF TIME SPAN AND TASK LOAD
ON PILOT MENTAL WORKLOAD

by

Scott L. Berg

and

Thomas B. Sheridan

ABSTRACT

Two sets of experiments were run to examine how a pilot's mental workload might be measured, and how these measures are affected by continuous manual-control activity versus discrete assigned mental tasks, including the length of time between receiving an assignment and executing it.

A fixed-base flight simulator was used, consisting of a Control Box, a high resolution CRT, and a PDP/11 computer. The Control Box contained a joy-stick, throttle, and all the switches and controls necessary for operating the simulator aircraft's electronic and mechanical systems. The Control Box inputs were fed to the PDP/11 computer. The computer used these inputs, the current state of the aircraft, and pre-programmed aircraft dynamics to update the aircraft's state and drive the CRT display. Aircraft dynamics were modeled on a Lockheed Jetstar business jet. The CRT display consisted of a forward, "out the window" perspective view and a cockpit instrument/indicator presentation.

The first experiment evaluated the strengths and weaknesses of measuring mental workload with an objective performance measure (altitude deviations) and five subjective ratings (Activity Level, Complexity, Difficulty, Stress, and Workload). Volunteer pilots flew a high intensity, manual-control mission and a high mental workload mission. Each mission type was flown over two different ground tracks. A method of activity analysis was developed for calculating relative mental and physical workloads and was found useful for like types of work, but unsuitable for directly comparing mental workload to physical workload.

In this experiment, overall subjective workloads were judged to be only moderate. Altitude deviations were greater for the high mental workload scenario although pilot subjective ratings were greater (more difficult) for the manual activity scenario. Mental workload appeared to reduce the pilots' ability to control their

altitude. Subjective ratings for the two scenarios were different, but their respective altitude deviations were similar.

The second set of experiments built upon the first set by increasing workload intensities and adding another performance measure: airspeed deviation. The pilots flew a low workload "Baseline" scenario, a high manual workload "Activity" scenario, a high mental workload "Planning" scenario, and a high manual and mental workload "Combined" scenario.

The degree of mental tasking had no impact on the magnitude of airspeed or altitude deviations. Five types of subjective ratings were elicited from the pilots. These proved different for the Activity scenario, less distinct for the Planning scenario, and almost indistinct for the very high workload Combined scenario. Relative to the Baseline scenario's subjective ratings, the incremental ratings for the Activity scenario plus those for the Planning scenario, equalled those for the Combined scenario. For the high manual workload scenario, all of the pilots gave similar subjective ratings. However, some pilots found the high mental workload scenario much more difficult than others did. Although altitude or airspeed deviations and subjective ratings did not correlate at moderate workloads, they did correlate at a high workload level.

The number of mental tasks had little impact on the percentage of mental tasks performed improperly. However, the level of manual activity had a decisive effect. High manual workloads resulted in a high mental task error percentage.

Although altitude deviations, airspeed deviations, and subjective ratings were similar for both low experience and high experience pilots, the low experience pilots had many more mental task errors.

The length of time from receiving a mental task to executing it had no effect on the likelihood that the task would be performed properly.

ACKNOWLEDGEMENTS

There are many people whose direct and indirect support made this project, and subsequent Master's thesis, possible. We'd like to thank Dr. Keiji Tanaka and Mr. Ahmet Buharali whose flight simulation and display graphics software were the foundation for the computer programs. In addition, Mr. Leonid Charny was a constant source of encouragement and computer advice. We would also like to thank the volunteer pilots who selflessly gave of their time, and all those unnamed others who helped at one point or another.

This work was supported by NASA's Ames Research Center under NASA Grant NAG 2-227.

TABLE OF CONTENTS

TABLE OF CONTENTS (cont)

TABLE OF CONTENTS (cont)

LIST OF TABLES

LIST OF TABLES (cont)

LIST OF FIGURES

LIST OF FIGURES (cont)

LIST OF SYMBOLS

ADI	Attitude Directional Indicator
AGL	Above ground level
ALT	Altitude above ground level (m)
ARTCC	Air Route Traffic Control Center
ATC	Air Traffic Control
Beta	Side slip angle (radians)
$\dot{\text{B}}\text{eta}$	Rate of change of side slip angle
CAT	Cockpit Activity Timeline
CRT	Cathode ray tube
CWS	Control Wheel Steering: a stability augmentation system
D	Distance to the runway
DCZ	Neutral stick position for pitch
DME	Distance Measuring Equipment; also, a distance measurement (nautical miles in the instrument display, meters within the simulation software)
ETA	Estimated time of arrival
ft	Feet
GREF	Reference glidepath, calculated by the autopilot as an intercept to the desired glideslope
GS	Glideslope
GSE	Glideslope error: deviation from the nominal glideslope
GSI	Glideslope Indicator: located next to the ADI, it displays aircraft vertical position relative to the glideslope
GØ	Simulator glideslope angle: -3°, -.05236 radians
H	Present altitude (meters MSL)
HC	Commanded altitude (meters MSL)
HSI	horizontal Situation Indicator
ILS	Instrument Landing System: a precision approach aid for azimuth and elevation guidance
kt	Knots: nautical miles per hour
LOC	Localizer: an instrument approach aid for azimuth guidance
m	Meters

MKS	System of measurement: meters, kilograms, seconds
nm	Nautical miles
p	Roll rate
psi	Present magnetic heading (radians)
psic	Commanded magnetic heading (radians)
Phi	Roll angle
PP	Physical Parameters: a type of workload measure
q	Pitch angle rate
r	Yaw angle rate
rms	Root-mean-square: a measure (feet)
SRS	Subjective Rating Scale: a type of workload measure
TPM	Task Precedence Map: a diagram of operator activities
u	Perturbed forward velocity
v	Airspeed (meters/second in the simulation software; knots on the instrument display)
VC	Commanded airspeed
VCRSE	Angular difference between a selected VOR course and the current aircraft magnetic heading (radians)
VDT	Video display terminal
VOR	VHF Omni-Range: a very high frequency radio navigation aid
VORE	Difference between a selected VOR course radial and the VOR radial at the present aircraft position (radians)
VVI	Vertical Velocity Indicator: an instrument which displays vertical velocity in feet per minute
WU	Workload Unit: a task measure covering a 30 second period
$\dot{z}$	Vertical Velocity (meters per second)
α	Alpha: angle of attack (radians)
δ_{ail}	Aileron deflection
δ_{col}	Pitch command
δ_{ct}	Throttle command
δ_{el}	Elevator deflection

LIST OF SYMBOLS (cont)

δ_{rud}	Rudder deflection
δ_{sp}	Spoiler deflection
δ_{th}	Throttle position
δ_{w}	Lateral stick command
θ	Pitch angle (radians)

Chapter 1

INTRODUCTION

My eight years and 2500 flight hours as a United States Air Force pilot kindled interests in aircraft cockpit design and the dangers of pilots operating near the limits of their mental and physical capabilities. The T-37, T-38, T-39, and B-52 aircraft which I had the privilege to fly, had cockpits designed in the late 1940's to the early 1960's. Over the years, aircraft modifications had resulted in some equipment and operating procedures which made a pilot's already demanding task even more difficult. The human factors community and the aerospace industry were aware of the problem and set out to use new technologies to lessen the pilot's workload.

Cockpit design practices of the last 15 years share a common thread: the degree and complexity of automation is increasing and accelerating. Current state-of-the-art designs such as the Boeing 757, 767, and Airbus Industries A310 have radically changed flight deck activities. Future designs, such as the U.S. Air Force's proposed Advanced Technology Fighter and the Navy's Advanced Combat Aircraft will demand far greater levels of automation because of the requirement to operate in an extremely hostile, changing environment.

Expert systems and artificial intelligence will reduce or eliminate certain types of pilot workload. However, in some instances they may simply <u>change</u> the type of workload. Pilots are operating less as manual controllers and more as supervisory controllers.

1.1 SUPERVISORY CONTROL AND MENTAL WORKLOAD

The "supervisory control" model of operator behavior describes the operator's role in planning, programming, monitoring, and intervening as necessary in some process [23]. For this portion of a pilot's workload, he monitors equipment and makes decisions.

The increased time and effort expended in monitoring aircraft equipment has raised concerns that in automating aircraft we may be raising the pilot's <u>mental</u> workload to unacceptable levels (or conversely, lowering it to undesirable levels). Thus, there is great interest in measuring this mental workload. However, measurement implies some level of understanding of the process. The degree in which one understands a process is often demonstrated by

the sophistication and accuracy of the "models" used to describe it.

1.2 MENTAL WORKLOAD MODELS

One widely accepted and useful model of the human operator was proposed by Jens Rasmussen [18]. His model (Figure 1) separated operator actions into three types of behavior: Skill-based behavior; Rule-based behavior; and Knowledge-based behavior.

Skill-based behavior pertains to conventional manual-control type tasks. The pilot combines his sensory inputs with his internal model of the aircraft's systems and certain rules or parameters to initiate some action. His senses supply feedback in a closed-loop control system to operate the aircraft's systems. Various optimal-control models have successfully predicted operator performance for these behaviors [13].

Rule-based behavior relates to various procedural activities such as deciding to lower the landing gear or initiate communications with Air Traffic Control (ATC). The pilot observes the state of his aircraft and its systems, associates those states with certain tasks, and decides upon some action based on his internally generated plans and stored rules. Fuzzy set models have been used to model this activity [26].

Knowledge-based behavior is the process of planning and making judgements. Using the information available to him and internal goals, the pilot plans how to perform the task. This plan is formed using rules pertaining to the task, and results in performing some action. This behavior is pictured as the outermost control loop, and typically embodies the slowest flow of information.

Jensen and Chappell [9], in their study on "Pilot Performance and Workload Assessment", found it necessary to modify Rasmussen's model. They felt that the Monitoring function was sufficiently different from Rasmussen's concepts of Rule-based or Skill-based activities to warrant designating it a separate category.

Sheridan and Simpson [24] used Rasmussen's model of the human operator to describe a pilot's task (Figure 2). Aircraft Systems and Environmental Factors such as turbulence, ATC requirements, and requests input into the pilot model. Within the pilot model, note that supervisory mental work is primarily a Rule-based process requiring short-term planning and memory (less than or equal to 60 seconds). However, the supervisory functions of planning and intervening relate to the higher level knowledge-based workload.

Other models deal with different aspects of the mental workload problem. Queuing theory is used to model the pilot as a discrete

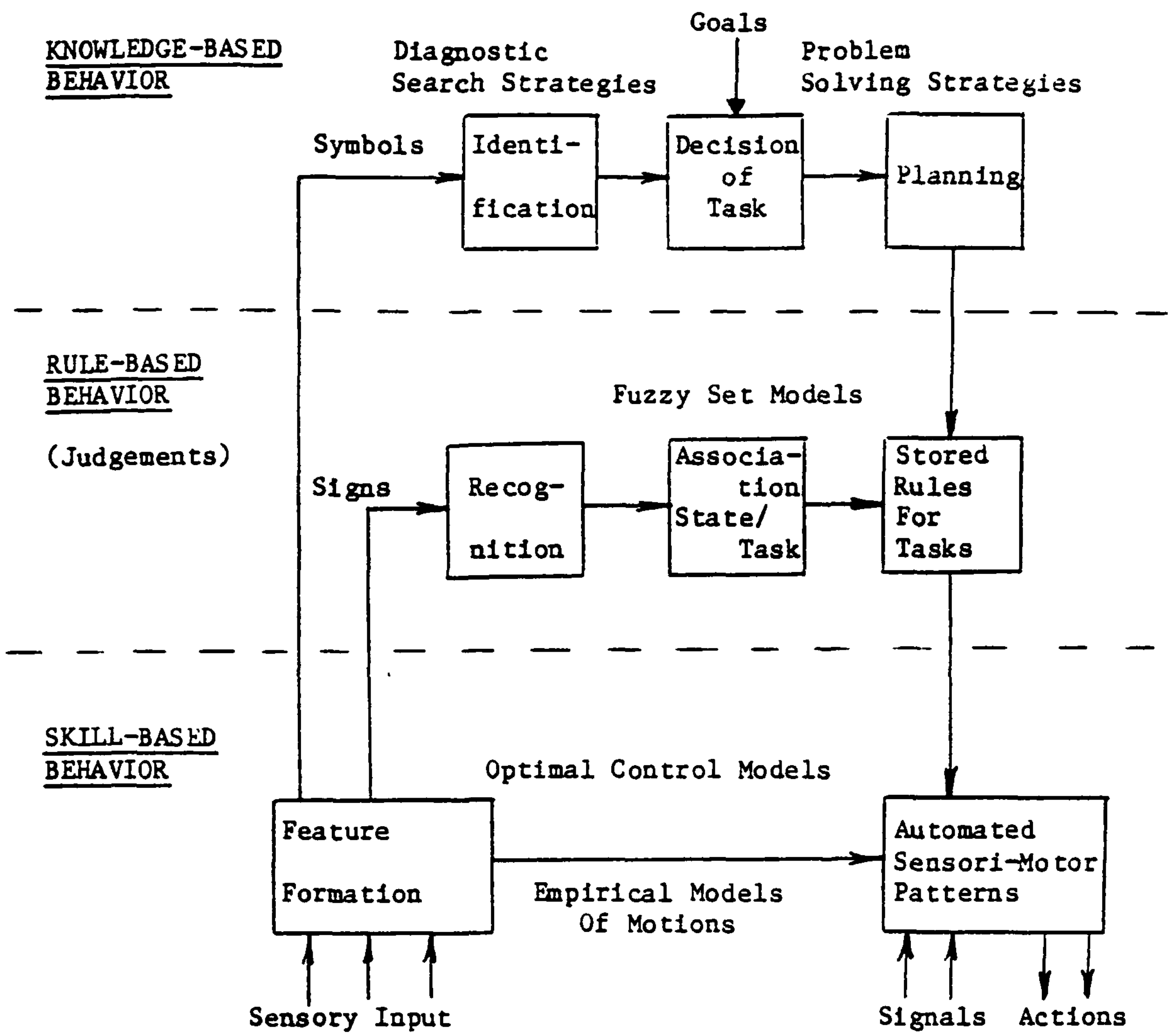


Figure 1: Rasmussen's Cognitive Model

ORIGINAL PAGE IS
OF POOR QUALITY

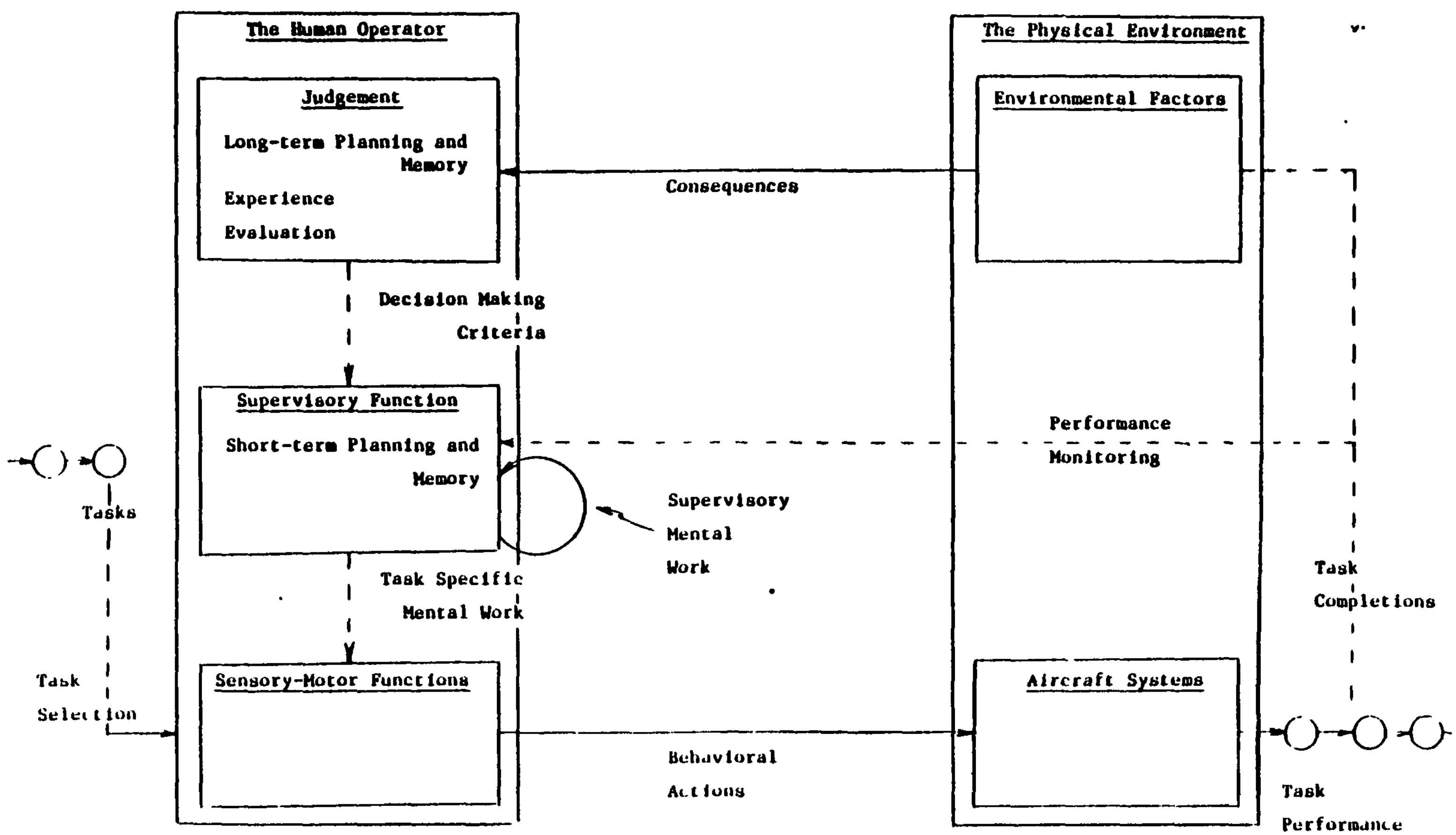


Figure 2: Sheridan and Simpson's Qualitative Paradigm for Pilot Mental Workload

data sampler who establishes several event queues to accomplish required tasks. Thus, the cockpit is a "multi-queue" environment and forces the pilot to rotate his concern from one task to another, allocating attention as necessary. When a pilot is busy, tasks begin to pile up, the queue lengthens, and performance theoretically degrades. This degradation occurs for several reasons. There are delays in accomplishing tasks. There is also an increased probability of task interruption due to the arrival of higher priority tasks. Or, some tasks may be omitted because queue size exceeds the pilot's short-term memory [3, 21, 28].

This limited short-term memory capacity of the human operator is directly addressed by models which describe the human as a limited capacity information channel. The fact that people have a limited memory capacity has been known for centuries. However, G. A. Miller [17] first put this fact into information theory terms in 1956. He pointed out that stimuli which varied from one another with respect to only one attribute, could consistently be assigned to no more than seven categories without error. Others have shown that information transmission rate is limited. Figure 3 illustrates one information channel model [22].

The limits of the human operator as an information channel have three important aspects. First, there are absolute limits to a person's capacity to both remember and transmit information. Forgetting, lack of understanding, and memory saturation result in a loss of information. Second, some parallel processing can be carried out for coordinated tasks, but to do several independent tasks requires switching among them. This requires multi-queue mental processing models. Third, when working at capacity, one can increase speed only at the expense of accuracy, and conversely.

A common prediction is that task performance will decline as mental workload increases beyond a certain point. In its most general form, predicted task performance is believed to be a function of mental workload, and can be pictured as a series of curves remarkably similar to a coefficient of lift versus angle of attack plot for airfoils (Figure 4). It is also commonly asserted that, as shown in Figure 4, increased operator skill results in increased performance at a given level of mental activity, or decreased mental work for a given performance level.

Sheridan and Simpson [24] theorized that when heavy workload forced a pilot to choose tasks and allocate his attention, "...the non-task-specific short-term planning or 'supervisory' component of mental work increases..." This increases the pilot's uncertainty, anxiety, and generalized stress. Under such circumstances, "...the pilot's skilled behavior will be compromised."

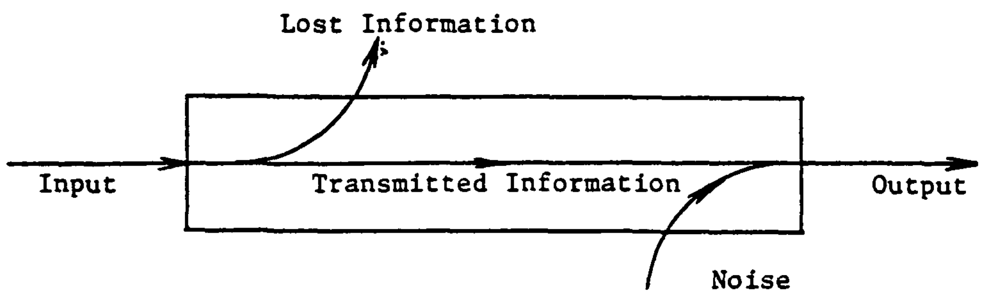


Figure 3: Model of an Information Channel

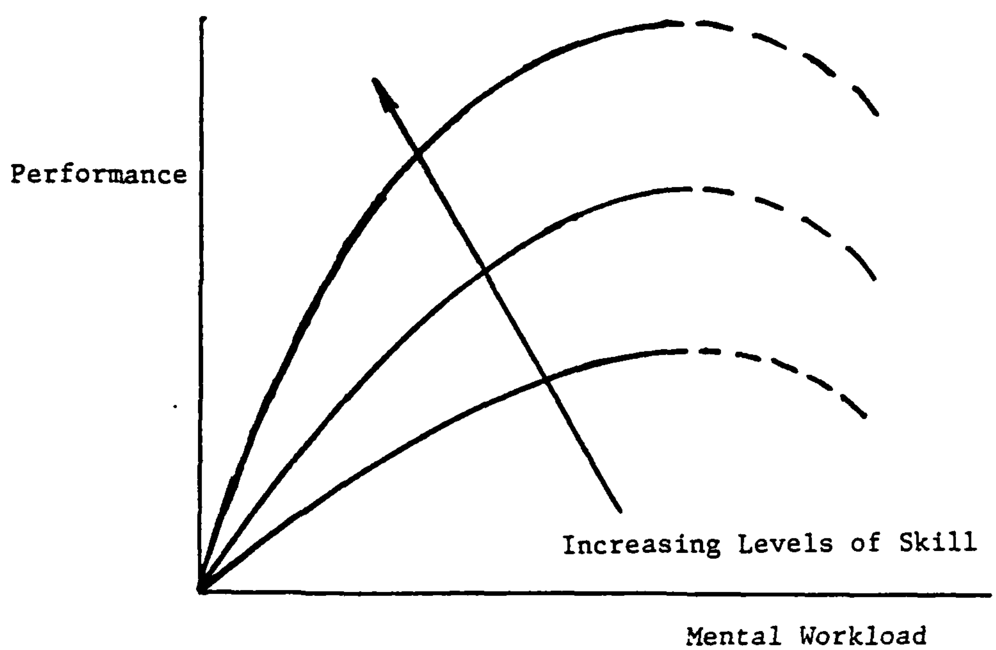


Figure 4: Performance as a Function of Mental Work

1.3 DEFINING AND PREDICTING MENTAL WORKLOAD

The greatest problem in trying to find ways to minimize mental workload is simply trying to measure or quantify it. Measuring physical workload is relatively straightforward. One can measure calories expended, numbers or rates of movements, forces exerted, pulse rate, blood pressure, et cetera. However, it is difficult to measure something which is poorly defined, and there has been disagreement over what constitutes mental workload.

After an extensive literature review, Williges and Wierwille [31] stated that there is no agreed upon definition of mental workload, nor single, universal metric of it. "Mental workload is a theoretical construct, and as such might best be defined operationally." Systems Engineers, Psychologists, and Physiologists all have their own methods of defining and measuring mental workload.

Given that there is disagreement about the definition of mental workload, there is a further problem: prediction. Predicting mental workload is important to the designer, but it is just as important to the investigator. The measurement of an unknown and poorly defined quantity may produce misleading results. Several predictive techniques have been used and are helpful, but are not definitive.

Cockpit Activity Timelines (CATs) are used extensively to quantify physical workload. CATs break down activities into discrete physical actions such as "reaching 6 inches over head", or "pressing button". However, not all cockpit tasks are identifiable or measurable. For example, how does one deal with "decide to request a change of flight plan from ARTCC"?

There is also some degree of arbitrariness in detail. One CAT may say, "lower landing gear". Another may say, "reach 18 inches forward and 6 inches left; grasp landing gear handle; lower handle; bring hand back to throttle; wait for landing gear warning light to go out, warning horn to silence, and hydraulic pressures to stabilize".

Finally, CATs are not as precise as they seem to be. Time of execution, for example, will vary with the individual pilot, the pilot's mood, instantaneous workload, et cetera. However, given all these drawbacks, CATs have been useful for rough estimates of physical effort requirements and are widely used.

Task Precedence Maps (TPMs) are also widely employed. TPMs are a schedule of events. They delineate the occurence of physical events and the beginning or end of some task as a function of time. They are most useful for a macroscopic analysis of activities.

One problem with using these techniques to predict mental workload was pointed out by Hart and Bortolussi: "the workload

associated with a complex task may be considerably different than would be predicted by combining the workloads of the component tasks." [5] Another problem is that mental workload is not simply a function of the aircraft or procedures; it is also a function of the pilot.

1.4 MENTAL TASK CHARACTERISTICS

Given these problems, what are the characteristics of mental tasks? Mental tasks will arrive at random times. There will be uncertainty associated with some tasks: for example, must, should, how, or can one do the task? Different tasks have different priorities. Finally, some tasks require a specific sequence of processes.

Sheridan and Simpson further classified tasks into categories. [24] There are non-deferrable or pre-emptive tasks. These are usually operating tasks requiring immediate action such as turning on a piece of equipment or manipulating flight controls. However, they may also be mental tasks such as responding to an ARTCC request for information. Next, there are tasks which can be deferred for a short period of time (less than 60 seconds). These relate to monitoring activities, such as a pilot's instrument scan. Finally, there are tasks which are deferrable for more than 60 seconds, which in turn involve planning tasks such as deciding when, how, or whether to take some future action.

1.5 MEASURING MENTAL WORKLOAD: SUBJECTIVE MEASURES

How can one measure mental workload given these problems and uncertainties? Subjective Rating Scales (SRSs) have been used successfully by a large number of researchers. Two major reasons for selecting a subjective scale are: (1) mental events are not directly measurable; and (2) a person may compensate for increasing workload demand by increasing effort, thereby holding "objective" performance constant.

In choosing a subjective system, Sheridan and Simpson [24] decided to modify an already existing SRS: the Cooper-Harper Scale. The Cooper-Harper Scale has been used for many years by test pilots to evaluate aircraft handling. It rates handling qualities on a ten-point scale from Uncontrollable to Acceptable-Satisfactory (good enough without improvement). Like Cooper-Harper, Sheridan and Simpson separated a ten-point scale into four divisions: (1) impossible; (2) unacceptable; (3) unsatisfactory, but acceptable; and (4) satisfactory. Divisions 2, 3, and 4 were further divided into three subdivisions.

As a separate scaling effort, they suggested three main attributes to mental workload: (1) task time constraints; (2) task uncertainty and complexity of planning; and (3) psychological stress. This resulted in a three-dimensional scale. Each dimension had its own ten point scale with similar divisions.

Concerning SRSs, Rehmann, Stein, and Rosenberg [20] reported that "...these measures are often sensitive and provide meaningful data to the investigator." An investigation by Casali and Wierwille [2] on the value of 16 different techniques for estimating the pilot workload imposed by communications, reported that a modified Cooper-Harper scale reliably discriminated between low- and high-workload scenarios and between low- and medium-workload scenarios.

However, SRSs have some weaknesses. Katz [12] found in his study on "Pilot Workload in the Air Transport Environment" that "Perceived workload is not equivalent to performance." In that study, performance was judged on the magnitude of glideslope and localizer deviations on a simulated ILS approach. In addition, Williges and Wierwille [31] pointed out several other problems. First, the subject may confuse mental workload with physical workload in making the evaluation. Or, the subject may not be aware of the degree of mental loading. Also, subjective ratings are a function of emotional state, experience, learning, and natural abilities (although objective measures also share these influences).

Finally, post-flight interviews and questionnaires have proven valuable when used for supportive information.

1.6 PHYSIOLOGICAL PARAMETER MEASUREMENTS

Physical Parameters (PPs) have also been measured in an attempt to quantify mental workload. Casali and Wierwille [2] found that changes in pupil diameter reliably reflected communications workload differences between low- and medium-workload and low- and high-workload senarios. Mostly however, physiological measurements have been only marginally effective or completely ineffective in determining mental workload. Eye blinks, eye fixations, respiration rate, mean heart rate, heart rate standard deviation, electroencephalograms, and pulse rate measurements have all been evaluated and found wanting as practical measures of mental workload. [2, 8]

1.7 OBJECTIVE PERFORMANCE MEASURES

An extremely diverse assortment of objective measurement techniques have been employed and evaluated. One technique measures spare mental capacity. It assumes that the operator is a limited-channel sampler and tries to measure the difference between the operator's total workload capacity and the capacity needed to perform a task.

Two mathematical models have been suggested for the human operator. The task component/time summation model is essentially a computer simulation of workload. The information-theoretic model quantifies workload in terms of bits/second. Unfortunately, there has been only limited validation for either method.

Single primary task measures have been used with some success, but they are generally insensitive at low workload levels. Multiple primary task measures seek to overcome this limitation and provide a more complete picture of behavior and performance. Wierwille and Gutmann [30] found that a "...multivariate analysis of several primary measures has been demonstrated superior to one measure."

Casali and Wierwille [2] had success using measurements of errors of omission and commission. Errors of omission were valuable for distinguishing bewtween low- and medium-workload and low- and high-workload scenarios. Errors of commision were useful for distinguishing medium- and high-workload and low- and high-workload scenarios.

Several secondary task measures have also been extensively investigated. The nonadaptive arithmetic/logic technique measures performance on an arithmetic/logic task done during "free time". However, it is intrusive, can modify primary task performance, measures average instead of peak workload, and has not been found a sensitive workload indicator. A nonadaptive secondary tracking task technique has also been tried, but exhibits the same problems as the arithmetic/logic technique. Time estimation has been used with some success. However, it is only a relative, not an absolute measure. Nevertheless, Kantowitz, Hart, and Bortolussi [11] found that it "is possible to use an objective secondary task as an index of pilot workload..." especially if a synchronous secondary task "...occurs less frequently but coincident with critical events."

Adaptive arithmetic/logic and adaptive tracking techniques have been investigated, but they are limited to laboratory use because of equipment and safety considerations.

An occlusion technique which systematically provides or denies the pilot given amounts of data has been tried by several investigators with some success. However, it is intrusive, raises safety concerns in a non-laboratory environment, and is not very

sensitive.

Although various objective workload techniques have been used for many years and been successful in measuring physical workload, Kantowitz, Hart, and Bortolussi [11] pointed out that it has been far more difficult to achieve a useful objective measure of pilot mental workload than to find a useful subjective rating scale. One reason is that there is a great deal of "noise" inherent in these measures. Jensen and Chappell [9] said that although skill-based activities are easy to measure, the measurements can be difficult to interpret. Operators (or pilots) will often induce small errors to act as test signals and thereby gain additional information on system performance (pilot acting as a closed-loop control system).

Furthermore, the definition of a "significant" deviation becomes important. There is the possibility that the pilot may recognize a deviation and correct it before it reaches the "significant" level. In a system with high inertia, this may allow significant errors to go undetected. Thus, the actual error rates might be much higher than the reported or measured rates.

In addition, there is an accumulator effect. The pilot can act like a workload accumulator, maintaining a given performance level by working harder as the difficulty level increases. Individuals also set an arbitrary "acceptable" level of performance based on their own utilities. This level is normally short of their capacity, allowing "slack" for random or unusual events. Thus, until they near their performance limit, they can maintain similar performance levels by simply working harder. (see Figure 4)

Finally, objective measures may be insensitive across persons. That is, two people may show similar performance although one may be working much harder.

1.8 COMBINED MEASURES OF PERFORMANCE

In their study of mental workload, Tanaka, Sheridan, and Buharali [25] examined some implications of Rasmussen's behavioral model. They hypothesized that since skill-based, rule-based, and knowledge-based behaviors were different processes, they should cause different kinds of mental workload. Similarly, Johannsen [8] pointed out that there is a general consensus that mental workload has behavioral, performance, physiological, and subjective aspects. The result is that trying to measure mental workload with one measure is similar to trying to measure a swimmer's total energy output by instrumenting one arm muscle.

Thus, a number of researchers have proposed using several measures simultaneously. As Williges and Wierwille put it, "Because

of the multidimensionality of workload, it appears unlikely that any single measure will ever suffice completely." [31] (also see Leplat [14])

This multi-measurement approach has been used quite successfully. In one instance, Hicks and Wierwille [8] compared a number of mental workload assessment procedures for a driving simulator and found that, "...primary task measures and (subjective) rating scale measures...should be used in assessing driver workload, particularly if it is of a psychomotor nature."

However, although there is a general consensus that multiple measures are useful, applying this technique has not been uniformly successful. Attempting to explain inconsistencies in previous work, Kantowitz, Hart, and Bortolussi [11] theorized that "Perhaps one reason that objective and subjective workload data are 'de-correlated' may be that average and peak measures are being compared inadvertently." They then went on to demonstrate that properly designed objective and subjective measurement techniques could show congruous results.

1.9 GENERAL CAVEATS

There is one overriding caveat for the researcher, designer, or engineer who examines mental workload or applies the results of studies. As Sheridan and Simpson [24] put it, "...in the real world the subjective utilities of high performance on certain tasks may be considerably different than those found in the safety of an aircraft simulator." Although this fact is important in measuring physical performance, its relevance to the mental workload case is multiplied several times over because of the nature of mental workload.

Investigators also must deal with another "noise" source in any attempt to measure or analyze mental workload. Pilot errors are often used as indicators of workload level. However, pilot errors also induce additional workload. Hart and Bortolussi [5] investigated this problem in 1983. They reported that "...pilot errors...can alter the nature of the tasks that the pilot actually performs so that the workload experienced is substantially different from the workload that was intentionally imposed." Two significant results of their study were: "...errors are considered to be a significant source of workload and stress by experienced pilots"; and "...the pilots felt that the impact of errors on subsequent performance is very negative."

1.10 PROBLEMS TO BE ADDRESSED

I have examined the uncertainties in the model and definition of mental workload, the difficulties in measuring this workload, and the problems inherent in performing this research in a laboratory. Given all the previous qualifiers, this study will address several issues.

First, can mental workload be measured in a consistent, sensitive, and meaningful way? This issue was the thrust of an initial set of experiments which are described in detail in Chapter 3.

Second, is there a time-sensitive element in the mental workload indigenous to the aircraft flight deck? This question was examined in a second set of experiments, discussed in detail in Chapters 4 and 5.

Chapter 2

EXPERIMENTAL SET UP

2.1 GENERAL CONFIGURATION AND EQUIPMENT

Figure 5 pictures the laboratory flight simulator environment for this project. The volunteer pilot subjects manipulate controls and switches on a control box while getting aircraft state information from a MEGATEK cathode ray tube (CRT) display. The MEGATEK displays flight instruments, aircraft and equipment configuration, and a forward perspective view. The investigator has his own video display terminal (VDT) and keyboard for controlling the system.

Figure 6 is a diagram of the information flow for the set up. The pilot gets his visual information from the MEGATEK CRT and manipulates controls and switches on the control box. The investigator gets program status information on his VDT and directs commands to the Computer via a Keyboard. Control Box signals are fed to a PDP/11 Computer. The Computer's simulation program (see Appendix 1) takes the present aircraft state information, Control Box inputs, and the investigator's Keyboard commands to determine aircraft dynamics and a new aircraft state. The information is used to update the MEGATEK and VDT displays.

The basic aircraft dynamics were developed over a 12 month period by Keiji Tanaka. A great deal of experimental trial and error went into making the simulator's response as close as possible to the response of an actual aircraft. A number of pilots came to the lab, flew the simulator, and evaluated its handling qualities. Eventually, the simulation fidelity was brought to a high level, including realistic stall characteristics. I further modified the aircraft dynamics to make the flight controls slightly less sensitive and to improve the simulator's landing characteristics.

The Computer stores all Control Box switch or control manipulations and stores aircraft state data every 10.0 seconds. This data can be displayed on the investigator's VDT or printed out on a Line Printer.

The MEGATEK CRT display is Shown in Figure 7. The upper portion of the display shows a simplified, forward "out the window" perspective of an airport and three runways. Below this is a set of instruments in the familiar "T" pattern. An Airspeed Indicator, Attitude Deviation Indicator (ADI) with Glideslope Deviation Indicator (GSI), and Altimeter comprise the top row. A Horizontal

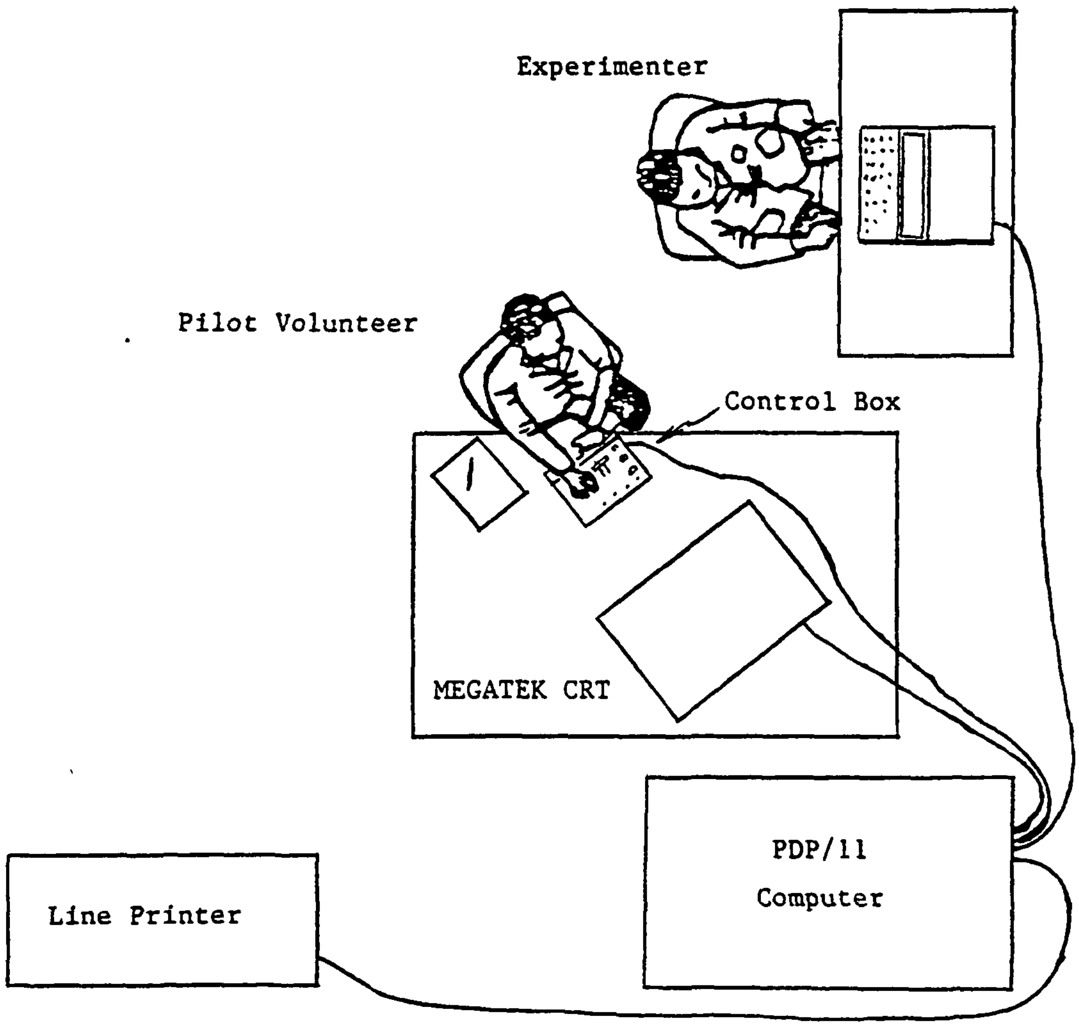


Figure 5: The Laboratory Environment

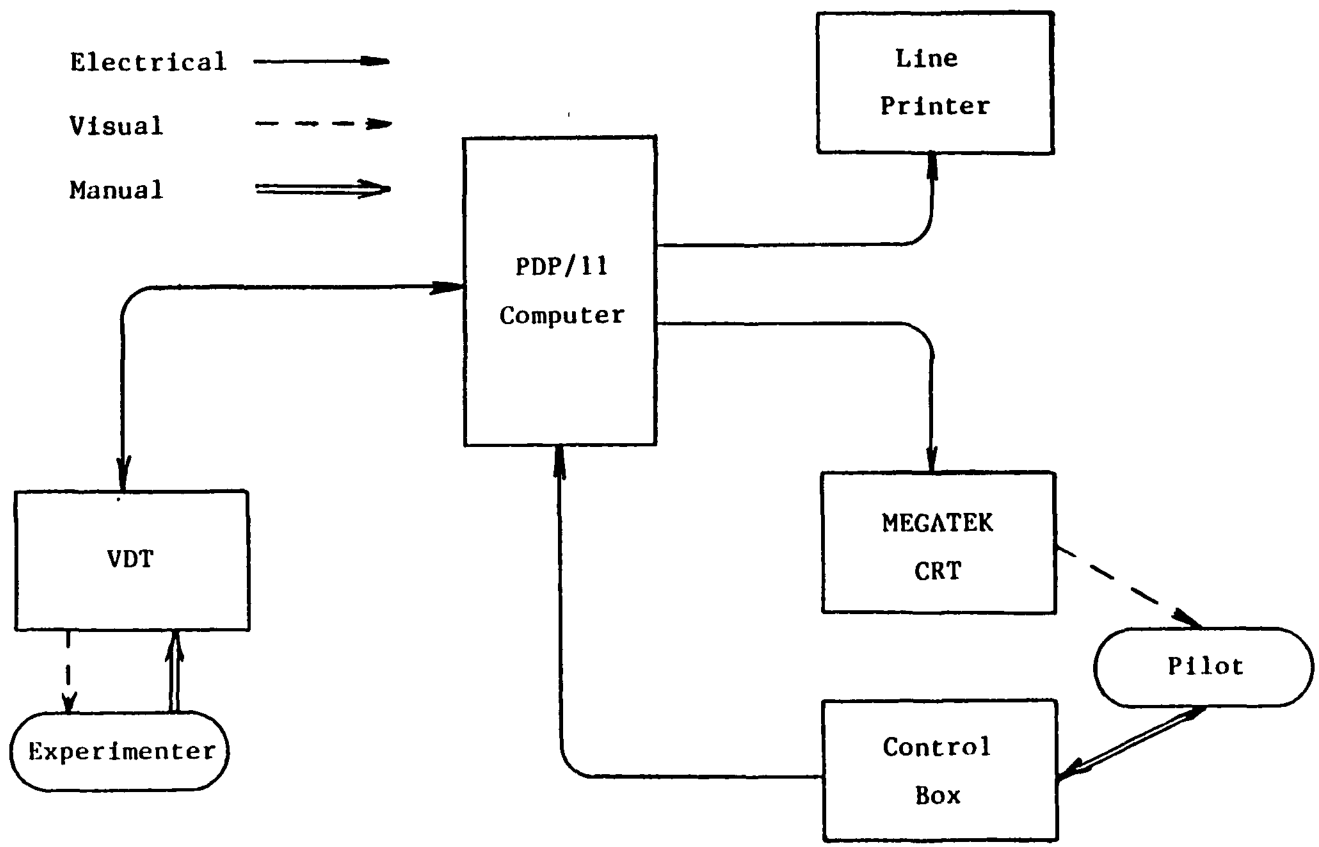


Figure 6: Signal Flow diagram of the Laboratory Set-up

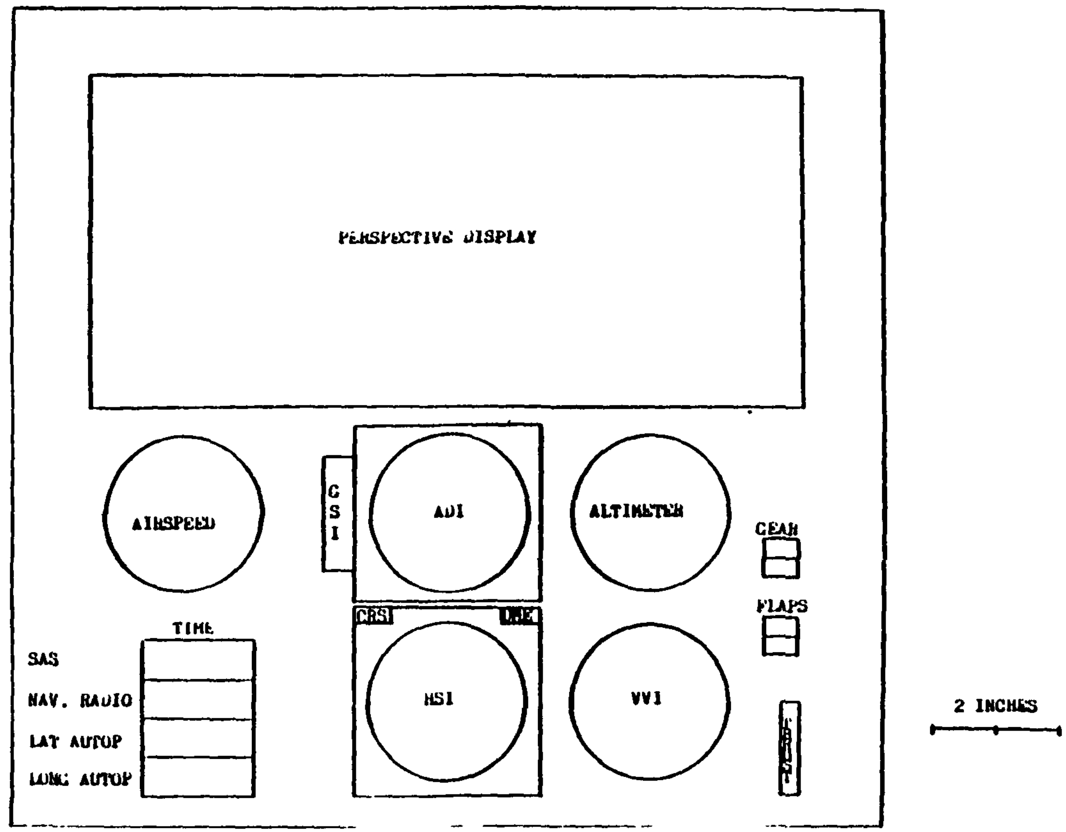


Figure 7: Diagram of the Simulator's high resolution MEGATEK CRT display

Situation Indicator (HSI) with the selected course (CRS) and distance (DME) to a selected navigation aid is directly beneath the ADI. A Verticle Velocity Indicator (VVI) is to the right of the HSI. Landing Gear Position (Up-Down), Flap Position (Up-Down), Thrust Setting, Stability Augmentation Selection (On-Off), Navigation Radio Selection (Off, VOR, ILS, channel number), Lateral Autopilot Selection (Off, Manual Heading, VOR Course, Localizer Course), and the Longitudinal Autopilot Selection (Off, Altitude Hold, Speed Hold, Altitude/Speed Hold, Glide Slope/Speed Hold) are also presented.

A drawing of the Control Box is shown in Figure 8. The subject interprets the flight information displayed on the MEGATEK and manipulates the controls and switches on the Control Box to make the "aircraft" respond in a desired fashion. The Control Box contains an aircraft-type control-stick or joy-stick, a throttle, and a number of other controls. On the top-rear of the box are eight Radio Toggles. To the left of the Throttle are the Course Set knob and the Flaps and Landing Gear Selector. To the right of the joy-stick is a longitudinal Trim Control. The front panel has six controls: Heading Set Knob; VOR/ILS Selector; Lateral Autopilot Selector; Longitudinal Autopilot Selector; Radio-Navigation Channel Selector; and Stability Augmentation Selector. For information on the lateral and longitudinal autopilot modes and the stability augmentation mode, see Appendix 2.

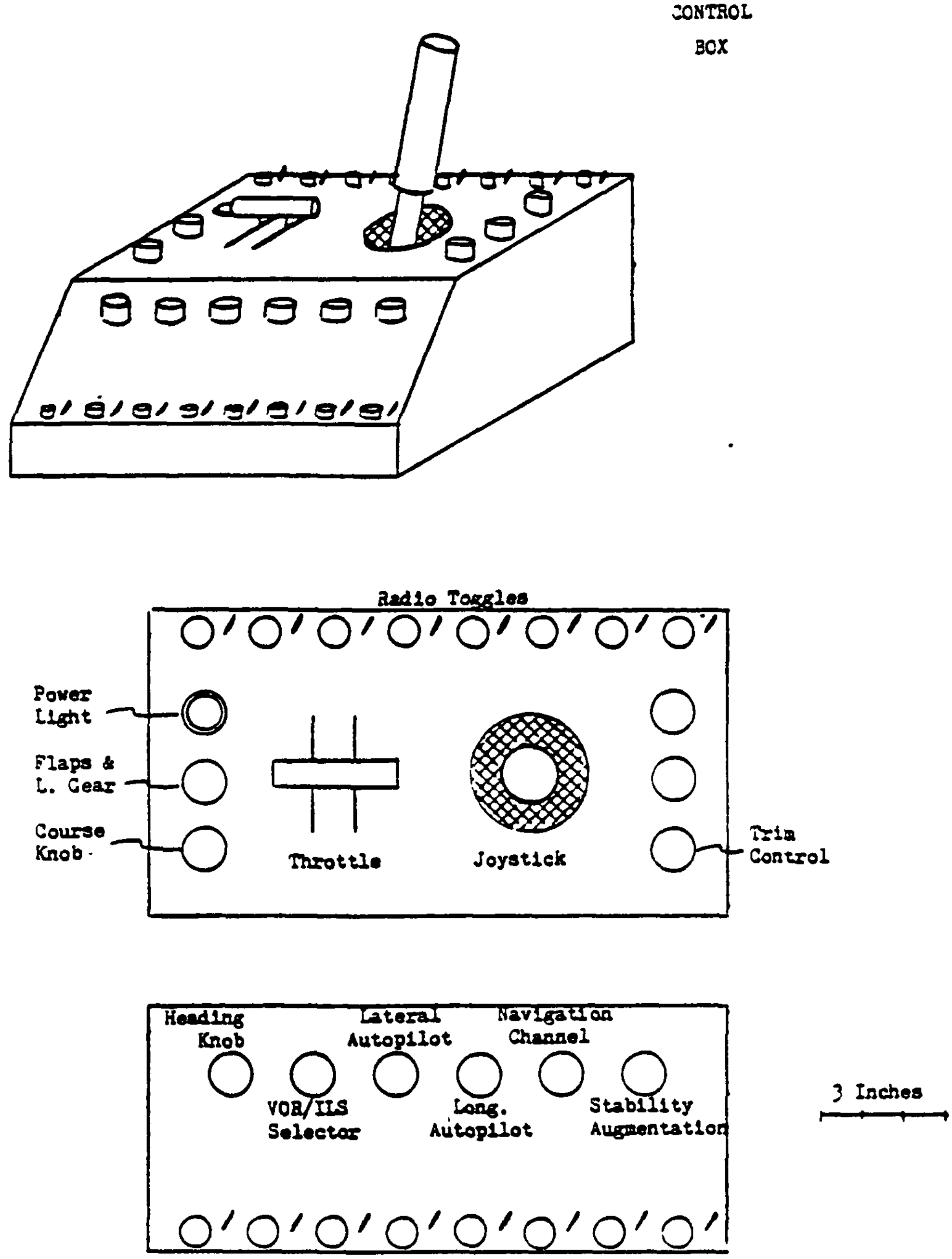


Figure 8: Layout of the Control Box for the flight simulator

Chapter 3

PRELIMINARY EXPERIMENT

3.1 SUBJECTS

Since these experiments demanded pilots who had, at a minimum, an instrument rating, I first recruited from M.I.T.'s resident military pilot population. Four very experienced pilots volunteered for the project. All four were Air Force officers. Two of the pilots had flown the simulator during previous experiments and the other two were given extensive training on the equipment before they were moved into the experimental phase.

The following is a summary of their flying experience:

A: Fighter-Type: ....1250 Hours
Jet: 1250
Total: 1250

B: Fighter-Type: 3200
Jet: 2750
Total: 3200

C: Light Aircraft:............ 550
Fighter-Type: 1000
Heavy Aircraft:............ 600
Jet: 1000
Total: 2150

D: Light Aircraft:............ 100
Fighter-Type: 700
Heavy Aircraft:............1300
Jet: 2000
Total: 2100

3.2 EXPERIMENTAL DESIGN

Two different ground tracks were used in this set of experiments. Figures 9 and 10 show the basic courses followed by the two routes: alpha and beta. Alpha was a clockwise route while beta was a counterclockwise route. Each pilot flew each route once per session. Two different routes were used in order to minimize the effects of transfering prior knowledge from one run to the next, "learning" the scenario, and consciously or subconsciously anticipating tasks.

Each route was flown in two versions. One version, labeled "Activity", was loaded with a number of tasks to perform. Most of these tasks were similar to the instruction, "Climb and maintain 4000". Such tasks exercise skill-based manual-control activity and short-term memory. These tasks exercise short-term memory because, in executing them, the pilot has to remember the particular altitude, heading, or airspeed he is trying to reach while he controls other parameters. The pilots were not allowed to use the autopilot as an aid at any point in these initial experiments.

The second, or "Memory" version exercised long-term memory by instructing the pilots to take some action at a given time in the future. An example of a long-term memory task is an instruction such as. "Descend to 2000 at Point Delta" given about 10 minutes prior to the aircraft arriving at Point Delta.

The two routes and two versions were counterbalanced between and within subjects. Table 1 shows the order in which the four subjects flew the four scenarios. Each subject flew only one session per day and each session contained two runs. Each session had runs exercising each of the two versions and each of the two routes.

Table 1: Order in which each pilot flew each scenario

	PILOT			
SCENARIO	A	B	C	D
Alpha Memory	1	2	3	4
Beta Memory	4	3	2	1
Alpha Activity	3	4	1	2
Beta Activity	2	1	4	3

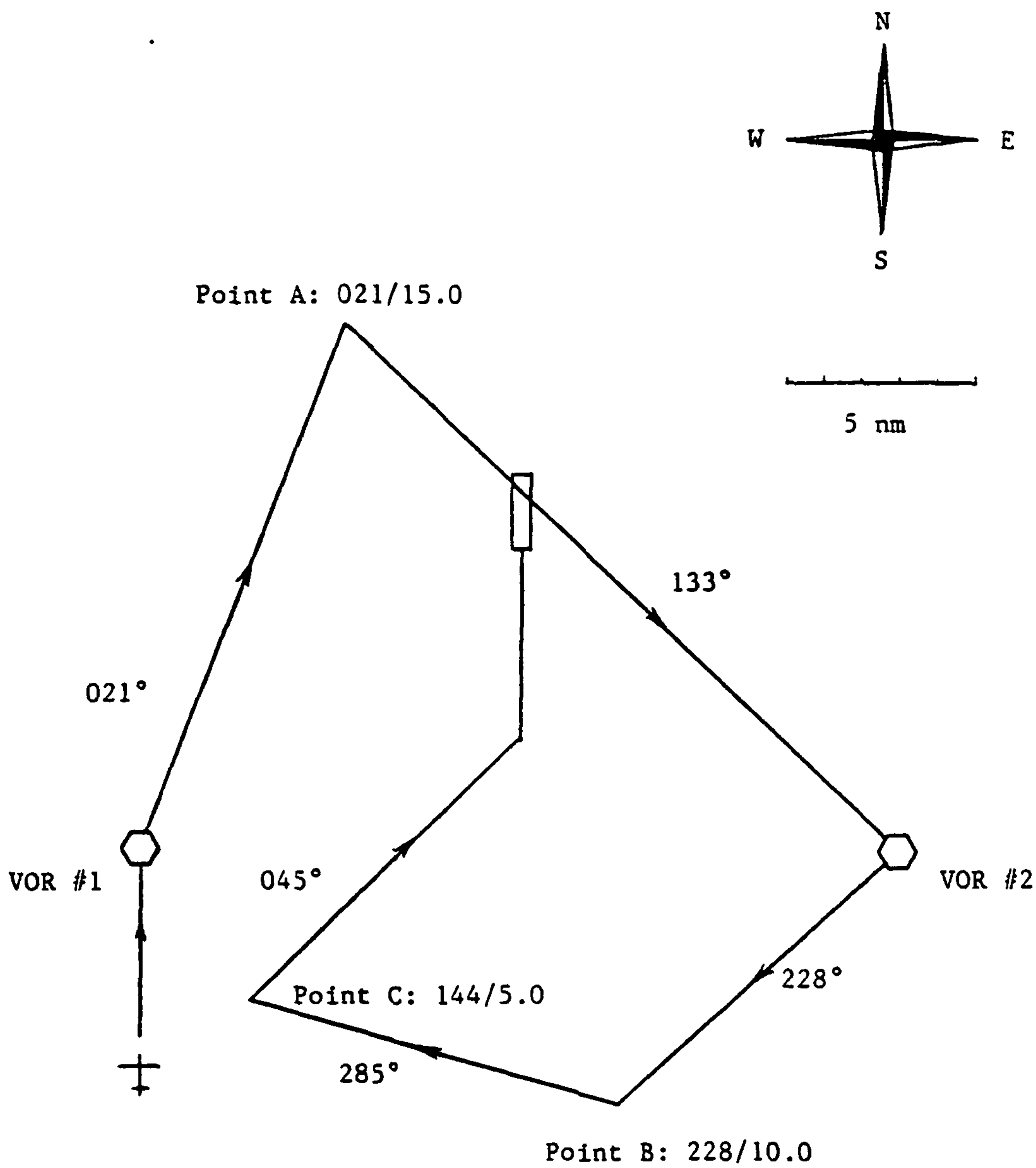


Figure 9: Alpha Route Navigational Chart

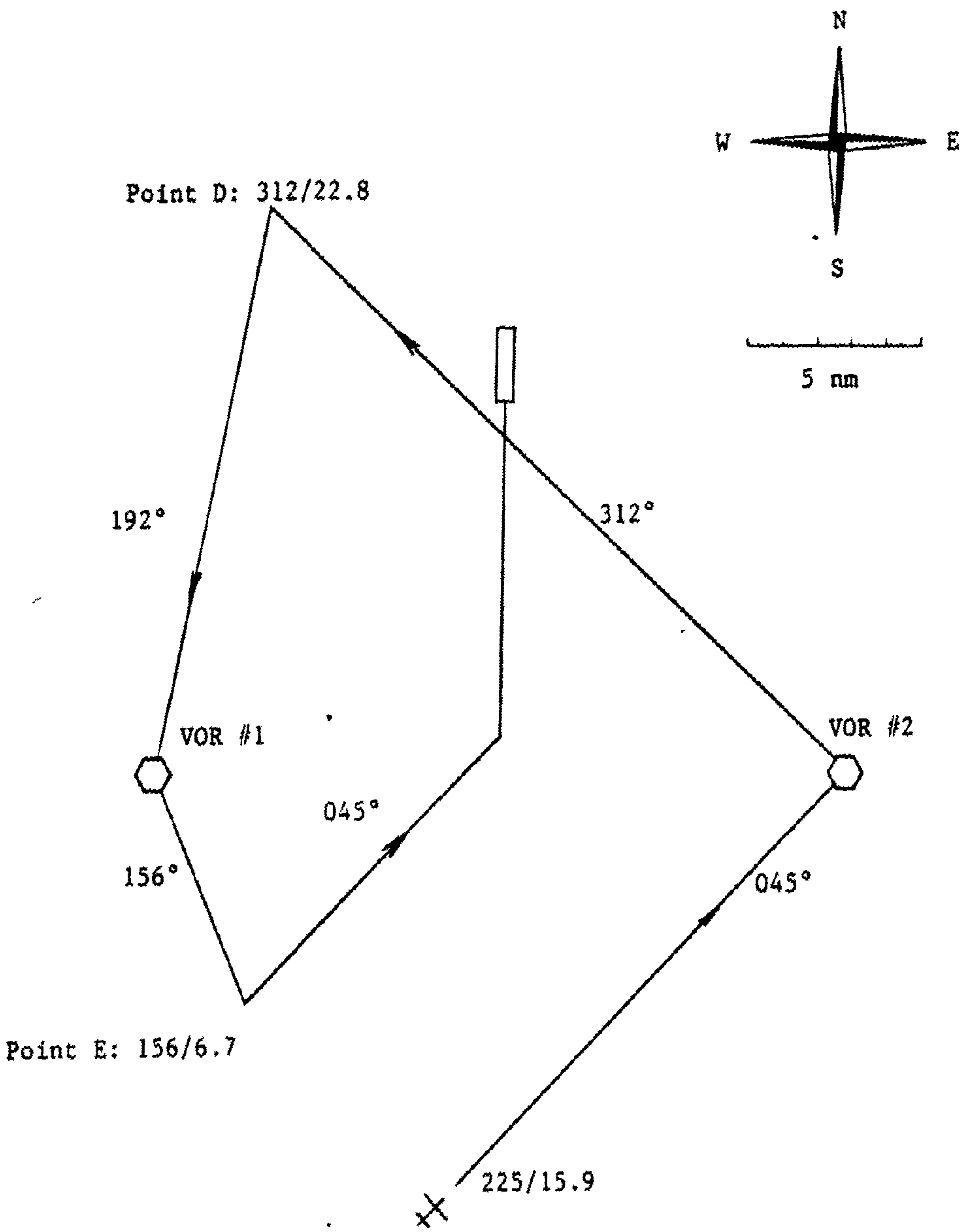


Figure 10: Beta Route Navigational Chart

"Navigation Charts" (see Figures 9 and 10) and note pads were provided to enable the pilots to record instructions (as in real flight). The Navigation Charts contained Navigation Aid positions, point identifiers, and the courses, bearings, and distances to and from various points.

Refering to Figure 9 for the alpha route, the pilots began by heading 360 degrees at 5000 feet, five nautical miles (nm) due south of VOR #1. After reaching VOR #1, they proceeded to Point A (VOR #1: 021/15.0), VOR #2, Point B (VOR #2: 228/10.0), and Point C (VOR #1: 144/5.0). The pilots then headed 045 degrees until intercepting the Localizer for an ILS to Runway 36 (ILS 4). The requirement to fly the entire route on instruments and perform point-to-point navigation, holding, and ILS approaches demanded a high level of pilot skill. (The "ceiling" was set at 1000 feet. Therefore, the perspective display showed nothing until the pilots "broke out" on short final.) The fact that the flights occured in a fairly small geographic area while flying at 200 $\pm$ 25 knots meant that at times things happened very quickly.

Figures 11, 12, 13, and 14 show the nominal ground tracks for the four scenarios. Figure 11 is the nominal ground track for the alpha route in its activity version. Note how ARTCC directed headings result in significant ground track deviations from a direct course. Figure 12 is the nominal ground track for the alpha route in its memory version. Note that there are few deviations, and thus, a much lower activity workload.

The differences between the task-loaded activity scenarios and the mentally-loaded memory scenarios is best illustrated by picturing the time histories of altitude, heading, and airspeed for each.

Figures 15 and 16 illustrate the time versus airspeed profiles for the alpha route. Note how the activity version has many more (10 to 2) airspeed changes. For the beta route, the ratio is 7 to 1.

Figures 17 and 18 document the number of heading changes for the two versions of the alpha route. The ratio of activity version to memory version heading changes is 13 to 7. The ratio for the beta route is 9 to 7.

Similarly, Figures 19 and 20 show the number of altitude changes for the alpha route's two versions. Alpha's activity scenario has 10 altitude changes while the memory version has 6 changes. For the beta route, the ratio is 10 to 5.

Every effort was made to make the total workloads of the alpha and beta routes as similar as possible while making the mental workload differences between the activity and memory versions as different as possible.

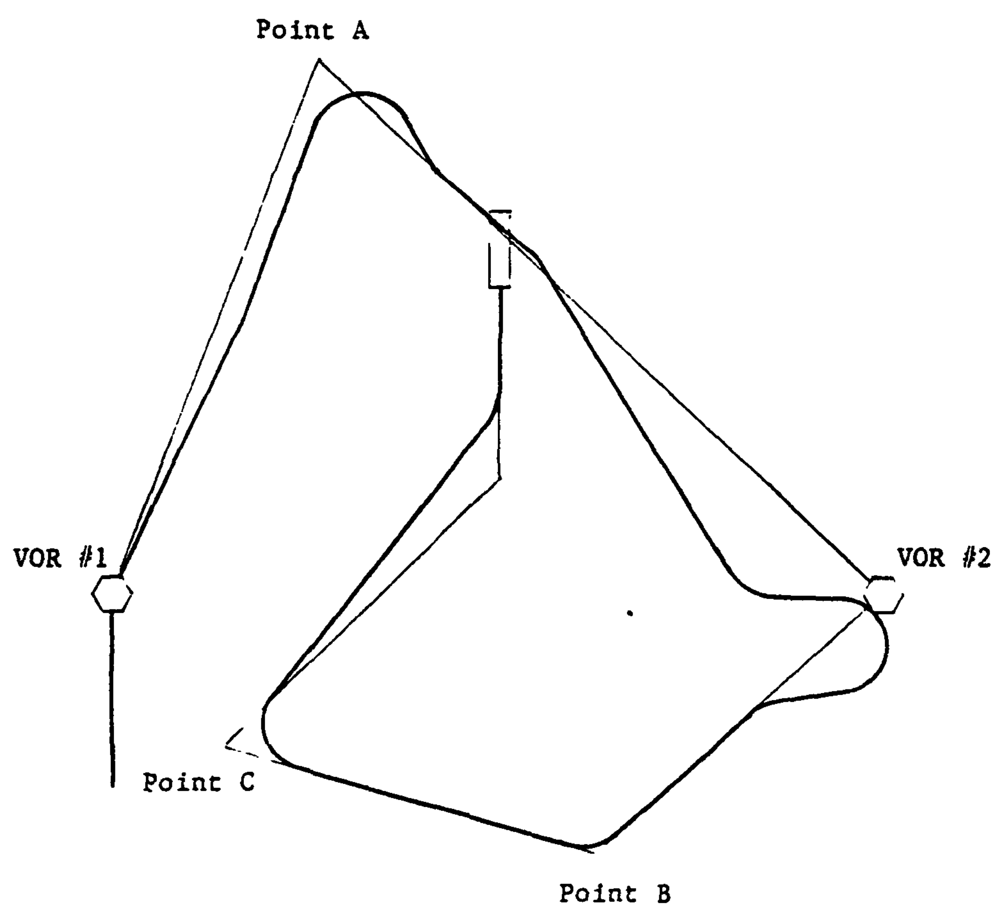


Figure 11: Nominal Ground Track: Alpha Route; Activity Version

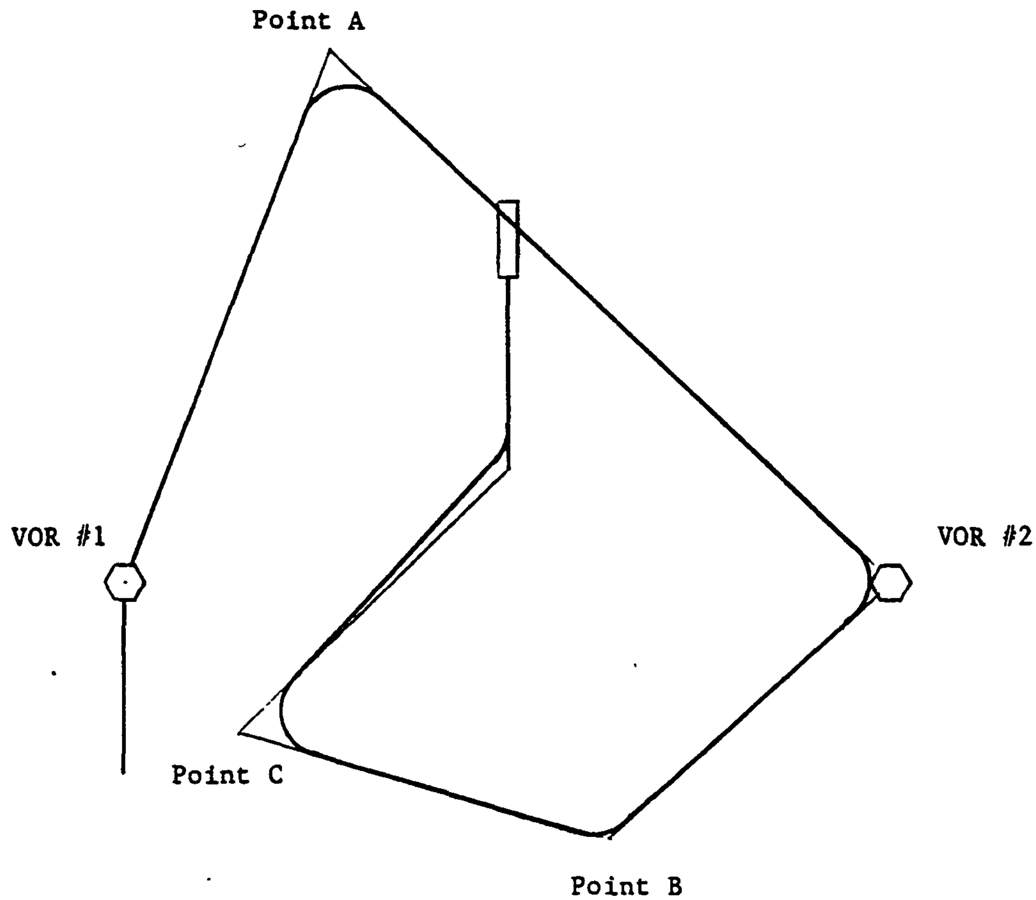


Figure 12: Nominal Ground Track: Alpha Route; Memory Version

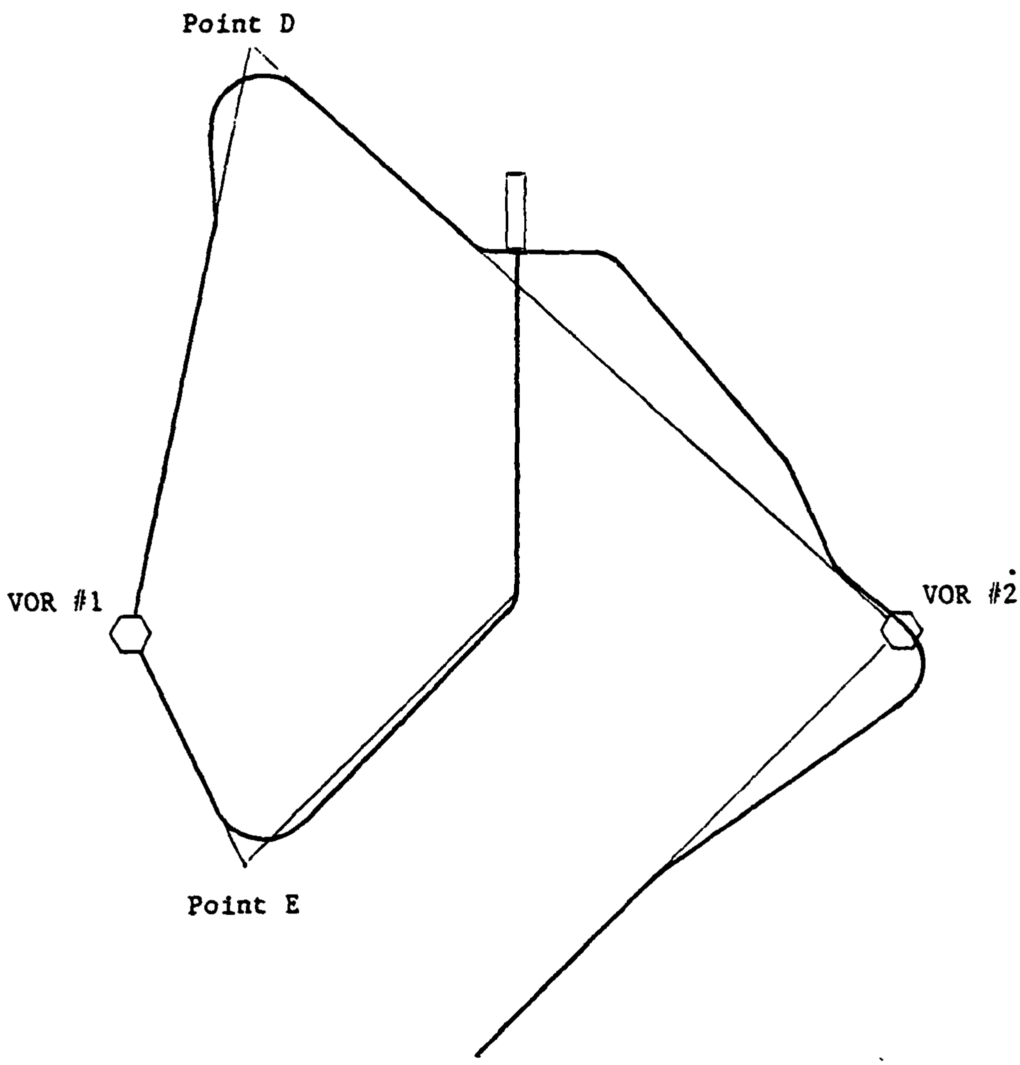


Figure 13: Nominal Ground Track: Beta Route; Activity Version

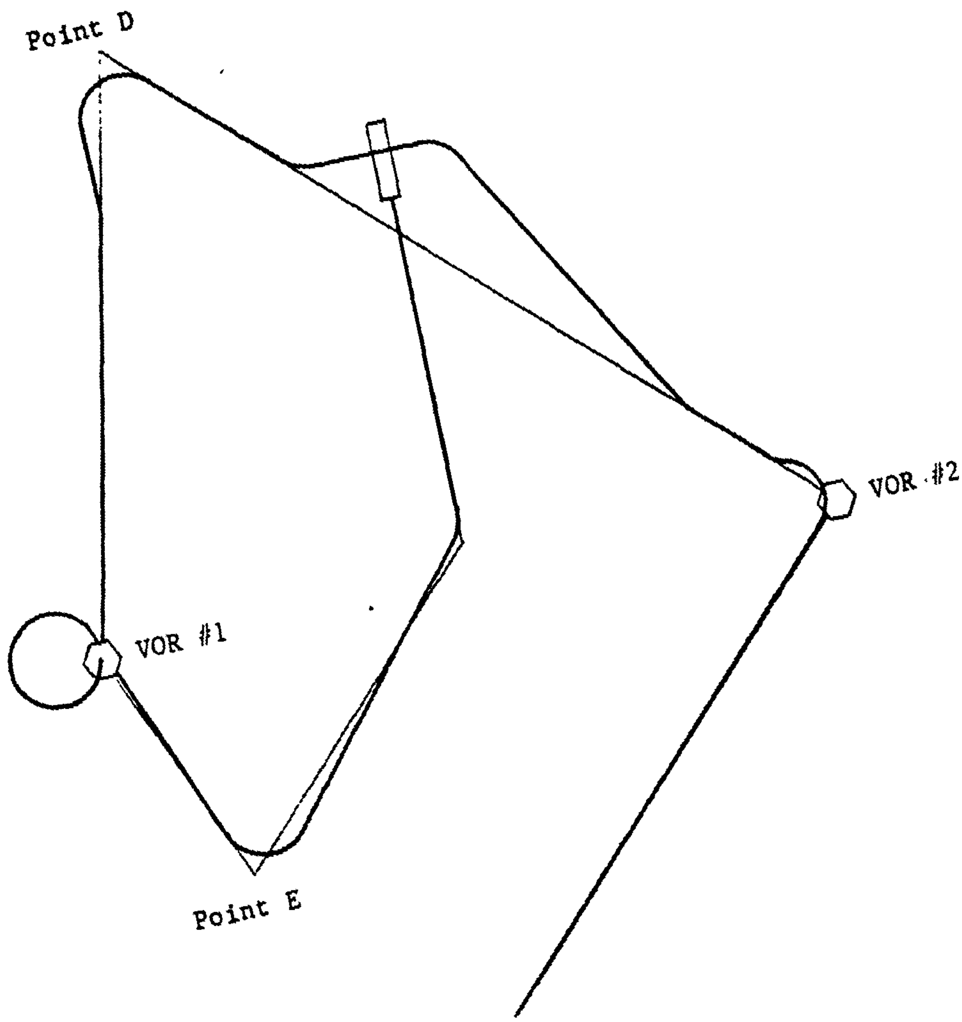


Figure 14: Nominal Ground Track: Beta Route; Memory Version

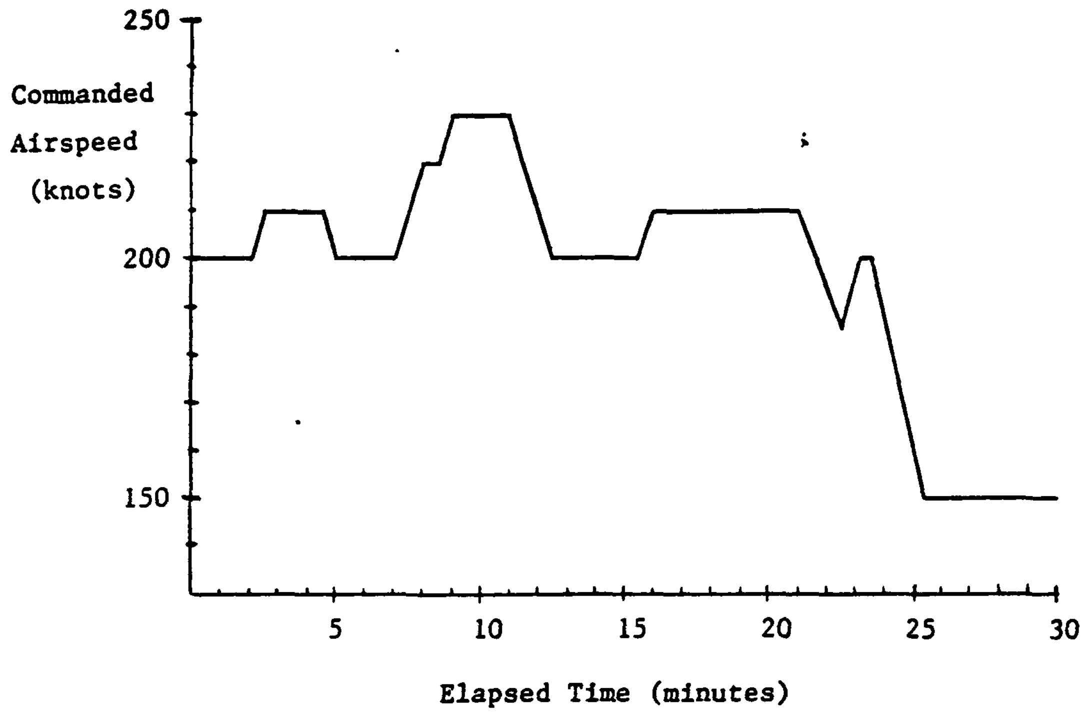


Figure 15: Commanded Airspeed for Alpha Route, Activity Version

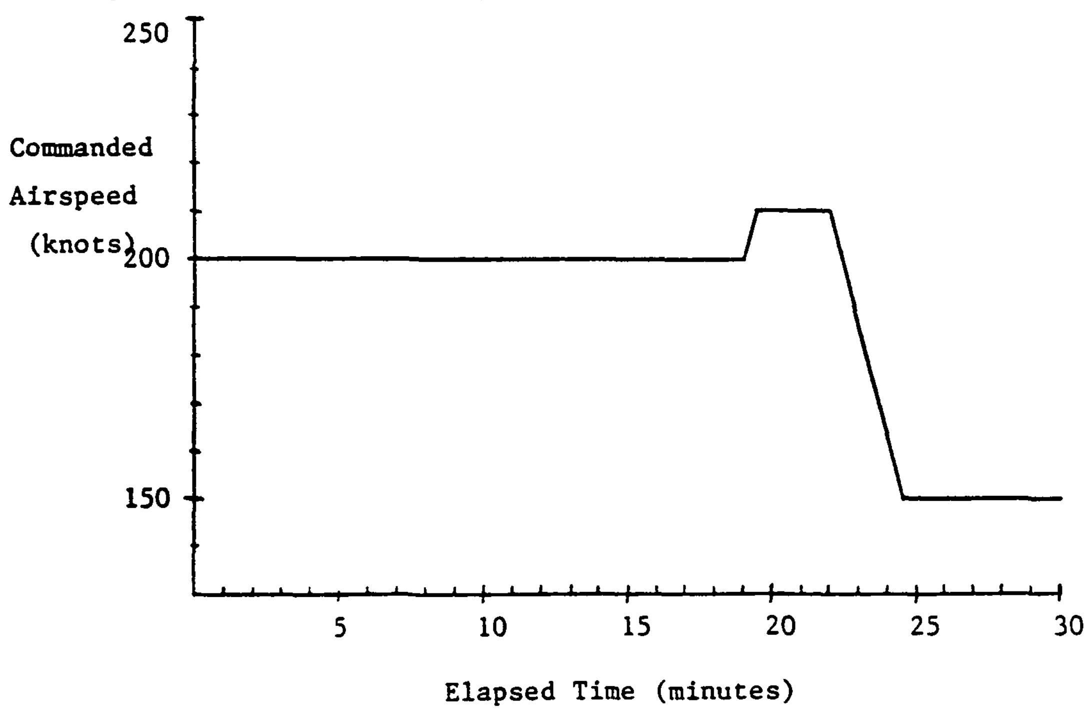


Figure 16: Commanded Airspeed for Alpha Route, Memory Version

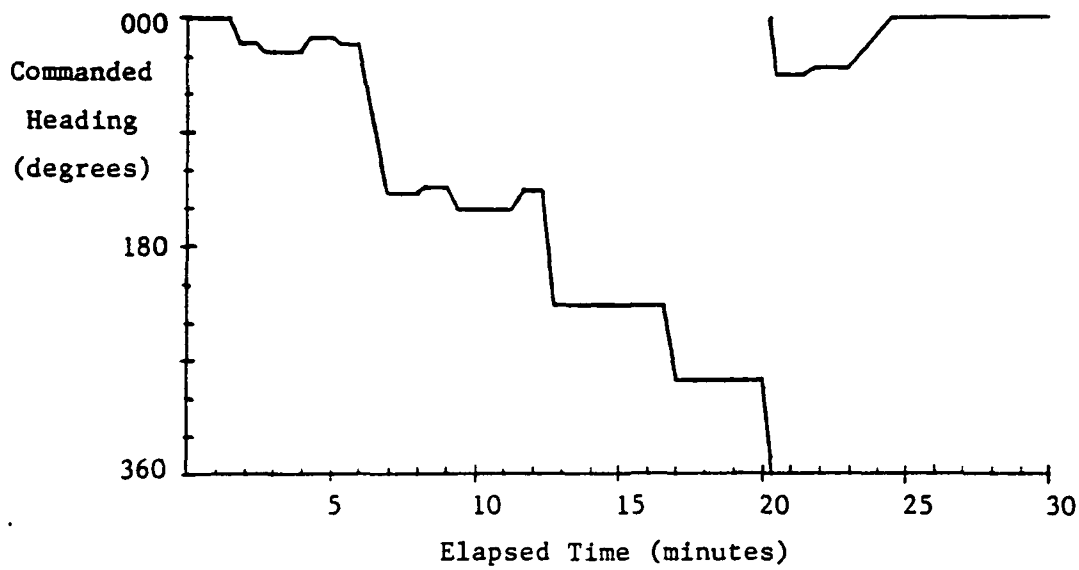


Figure 17: Commanded Heading for Alpha Route, Activity Version

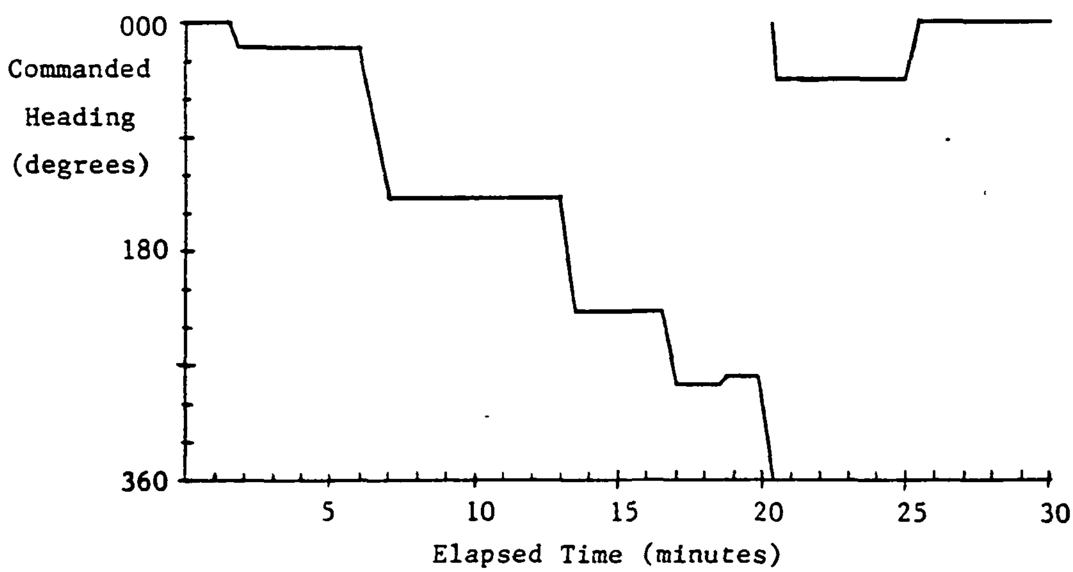


Figure 18: Commanded Heading for Alpha Route, Memory Version

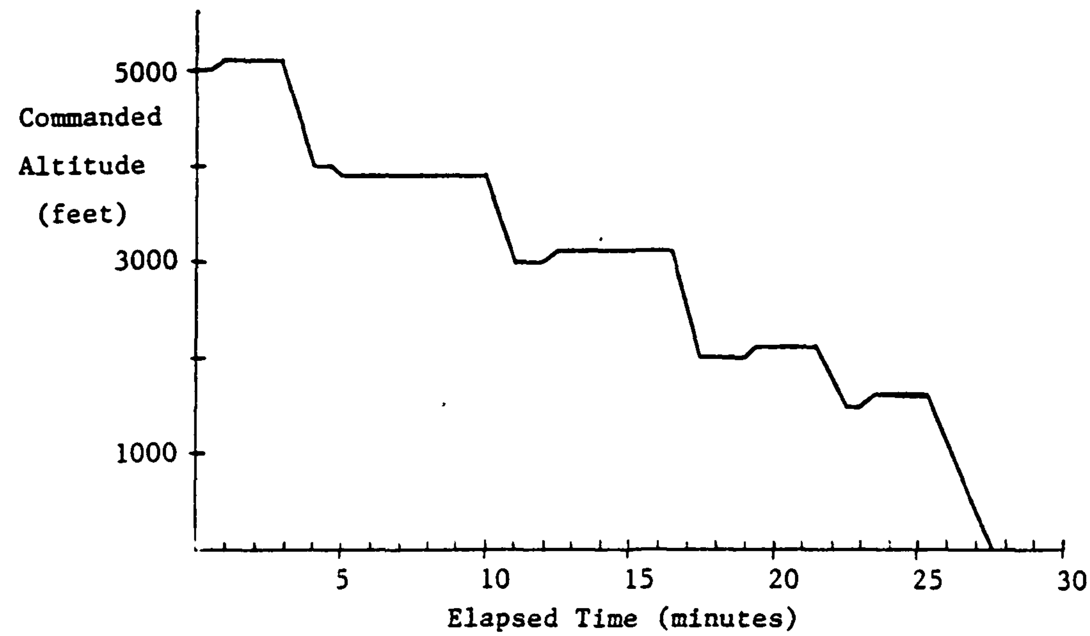


Figure 19: Commanded Altitude for Alpha Route, Activity Version

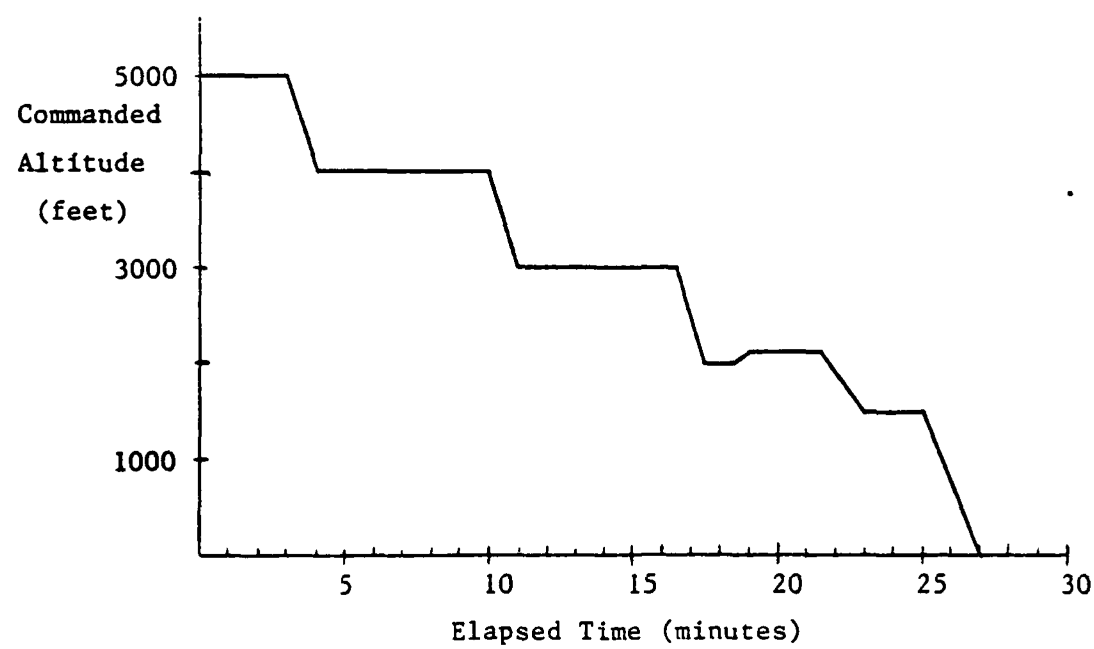


Figure 20: Commanded Altitude for Alpha Route, Memory Version

In an attempt to quantify the relative mental and physical workloads, I decided to use a type of modified Cockpit Activity Timeline. First, I calculated hypothetical mental and physical Workload Unit (WU) histories for typical activities. I divided time into 30 second blocks for these analyses.

For example, consider the activity "Climb 1000 feet". A pilot must make note of the request and inform ARTCC that he is initiating the desired action. I assumed that this would take about 15 seconds. Then, the pilot must climb 1000 feet. I used 1000 feet per minute as an average baseline for climbs and descents. Finally, the pilot was allotted 15 seconds for leveling off and making a level off report. Thus, the entire process took 90 seconds and this action was assigned three 30 second activity WU's.

In the process of performing this task, this activity was held in the pilot's short-term memory queue for the 90 seconds required for it. So, the task was defined as a short-term memory task and assigned three memory WU's.

For a long-term memory task, assume ARTCC directs "Report at Point Delta". The pilot must register the request, confirm it with ARTCC, and make some note of the requirement. This was worth one activity WU. When reaching Point Delta, the pilot had to contact ARTCC and make the required position report. This was assigned one activity WU. Therefore, the task generated one activity WU at the time it was directed, and one activity WU at the time of execution.

When the pilot receives the request, he places it in a long-term memory queue for monitoring over time. One hopes he doesn't forget the task, but retains it in memory. Thus, this task is given a series of mental WU's for each 30 second period between receiving the request and fulfilling it. (This method begs the question of whether memory actually functions in this manner. But, I felt that this method would be useful for measuring relative mental workload even if it did not accurately reflect absolute mental workload levels.)

Each of the four scenarios was broken down into a series of activity and memory tasks. Then, the WU time histories for these tasks were combined to produce plots for mental WU's versus time and activity WU's versus time. Figures 21 and 22 are plots derived for both versions of the alpha route.

These plots are not too enlightening, but they were useful. I used them to sum workload units over time for the four scenarios, and here differences were much more apparent.

Figure 23 shows the accumulated number of activity WU's as a function of time. This graph shows that the physical workloads were roughly equivalent for both routes' activity scenarios and both

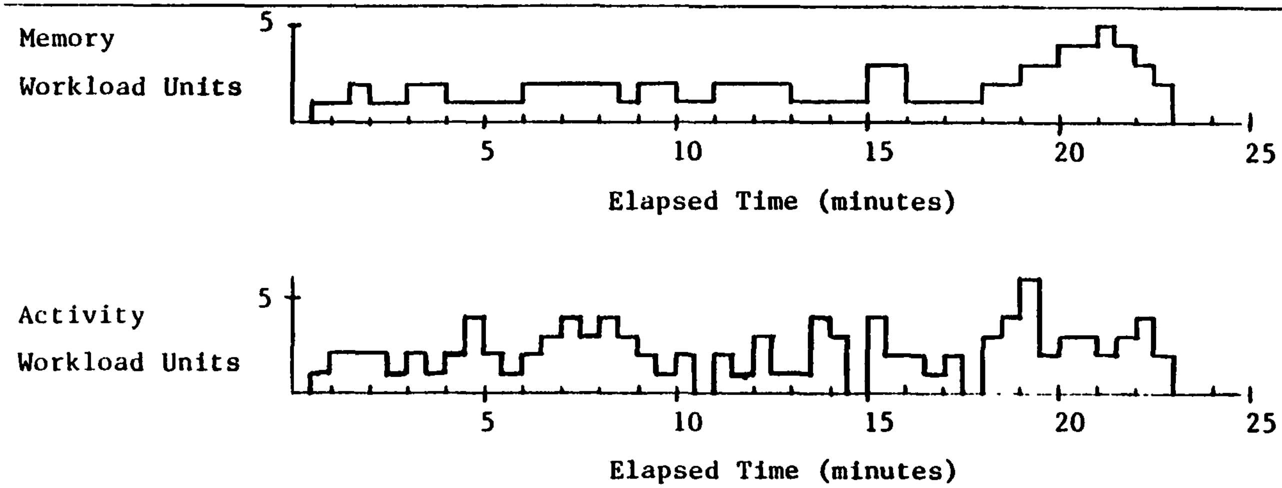


Figure 21: Number of Workload Units for the Alpha Route, Activity Version

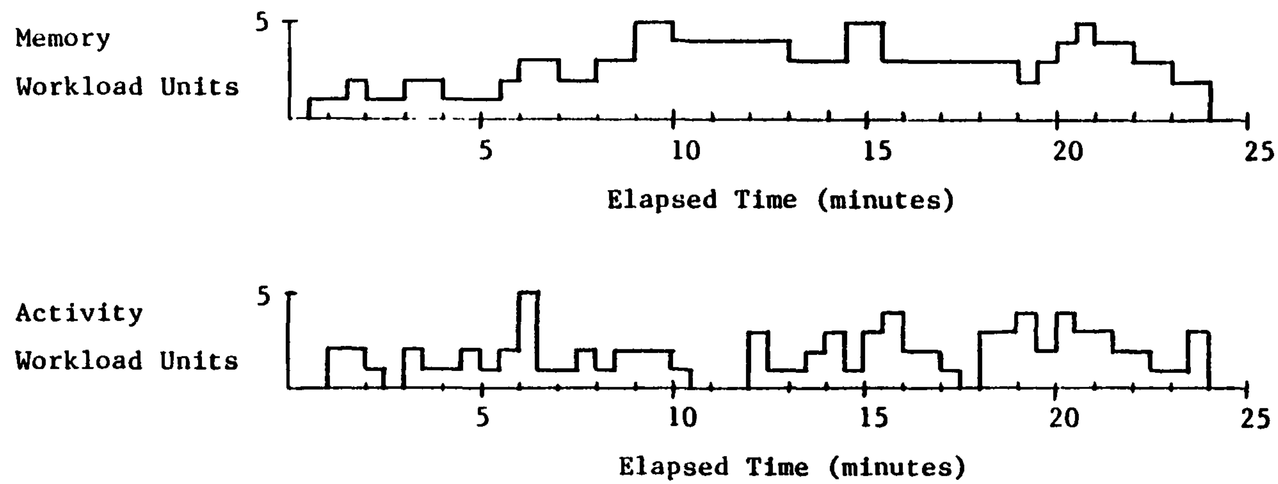


Figure 22: Number of Workload Units for the Alpha Route, Memory Version

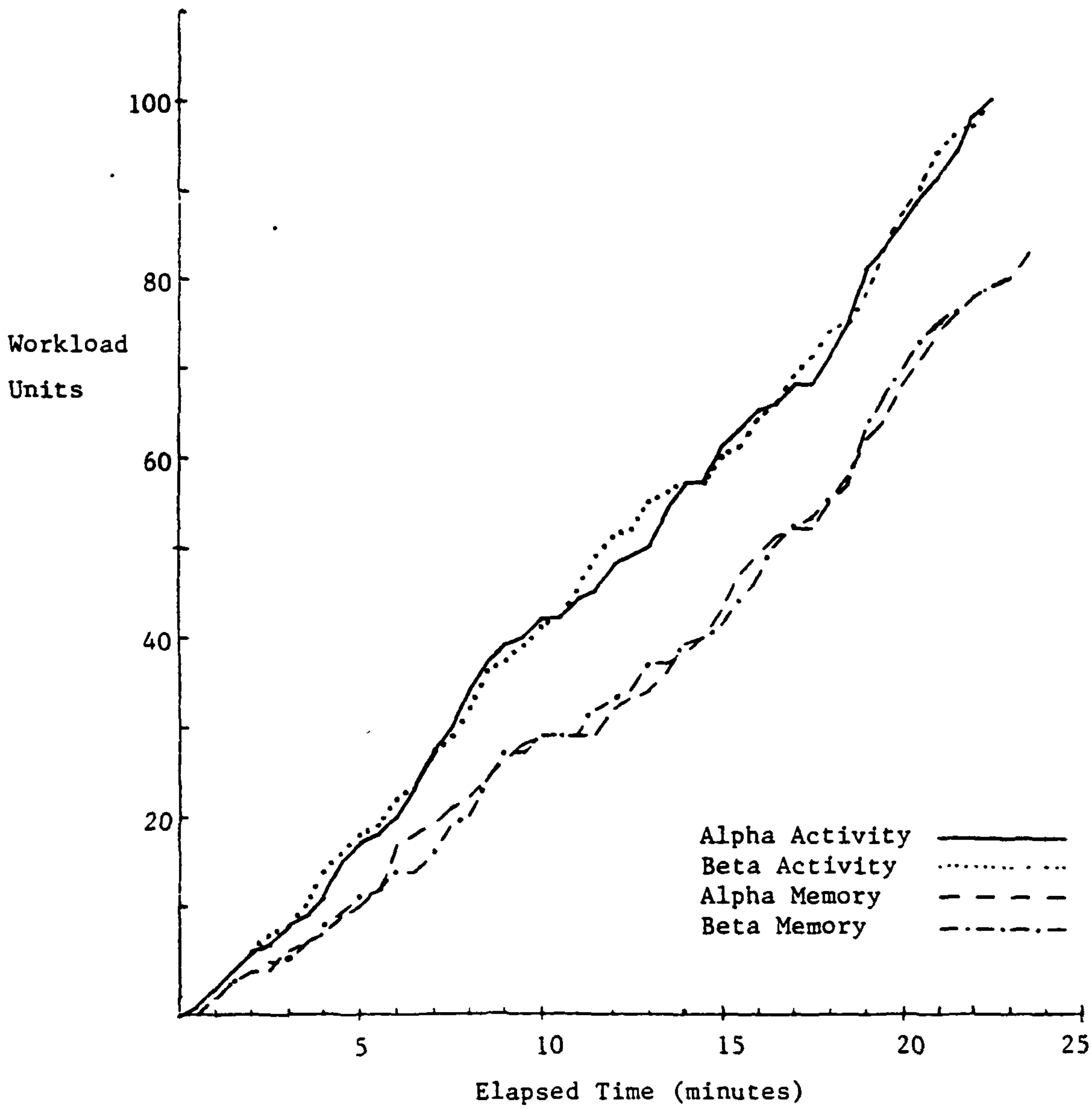


Figure 23: Accumulated Activity Workload Units

routes' memory scenarios. Also, the workload rate was similar for both activity scenarios and both memory scenarios. Finally, note the difference between the number of activity WU's for the memory versus the activity scenarios.

Figure 24 is a similar plot, but shows memory WU's instead of activity WU's. The same general comments apply as in the previous paragraph, but note how here the memory scenarios have the higher workload.

Figure 25 is a plot of the accumulated number of long-term memory tasks over time for each scenario. Note that the memory scenarios have roughly double the number of tasks as the activity scenarios. Again, the plots for alpha and beta routes are similar, and task rates are similar.

Figure 26 is a plot of the total number of memory tasks for each scenario. Note that the total number and relative rates of memory tasks were comparable for all four scenarios. However, keeping in mind Figure 25, the activity scenarios had a higher number of short-term memory tasks than the memory scenarios. These short-term memory tasks were mainly associated with the many activities within each scenario.

3.3 TRAINING AND INSTRUCTIONS

Before each session's data runs began, the volunteers spent 20 to 30 minutes flying the simulator. This practice consisted of changing headings, altitudes and airspeeds, intercepting courses, and making several ILS approaches.

When the pilots said they were ready and this investigator agreed that their performance appeared to have stabilized, they were given "Navigational Charts" (Figures 9 and 10) to study and the charts were fully explained to them. After any questions were answered, each pilot was given a page of instructions.

Figure 27 is a reproduction of the instruction sheet given to each subject. A few points deserve emphasis or explanation. The pilots were instructed to fly as "precisely" as possible. Thus, they were not told which aspect of their performance was being scored. They had to assume that <u>any</u> deviation might count against their performance. In addition, all simulated ARTCC instructions were handled verbally between the subjects and experimenter.

In addition to the instructions, each pilot was given a Subjective Rating Sheet (Figure 28) and a reference sheet (Figure 29) which explained "Workload" levels. The subjects were instructed to consider each scale as continuous and to regard the subdivisions solely as reference marks. A rating sheet was used for one day's

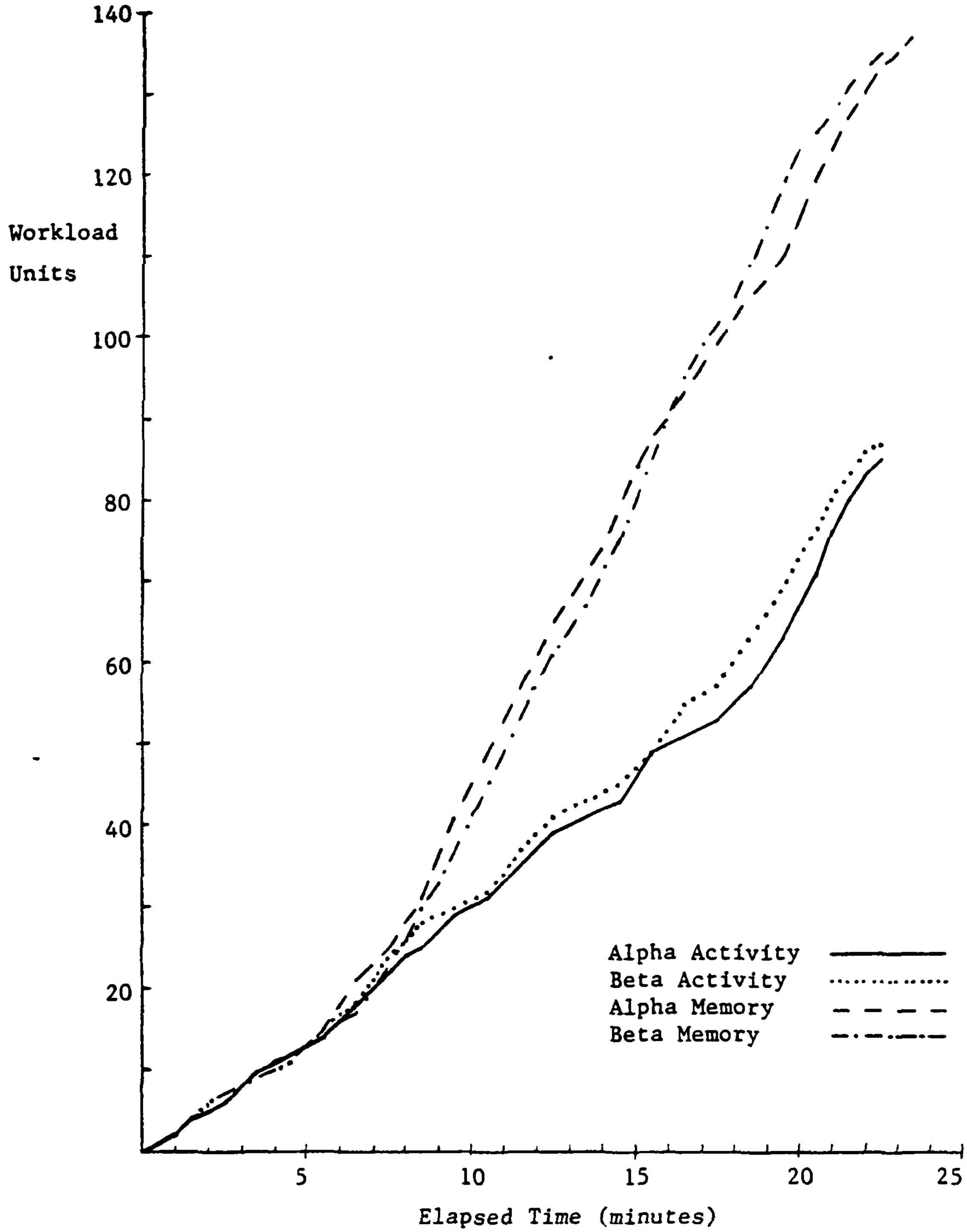


Figure 24: Accumulated Memory Workload Units

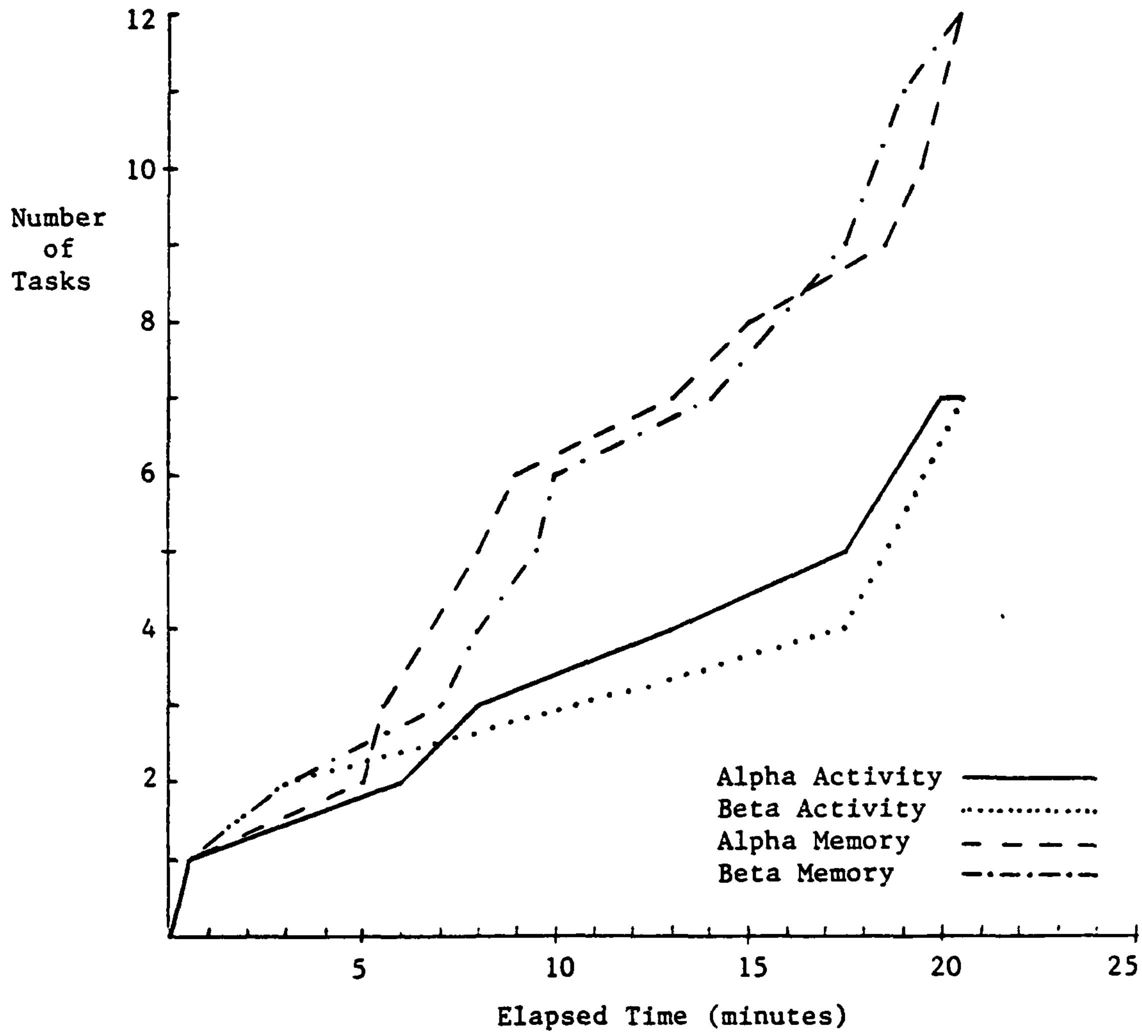


Figure 25: Accumulated Number of Long-term Memory Tasks

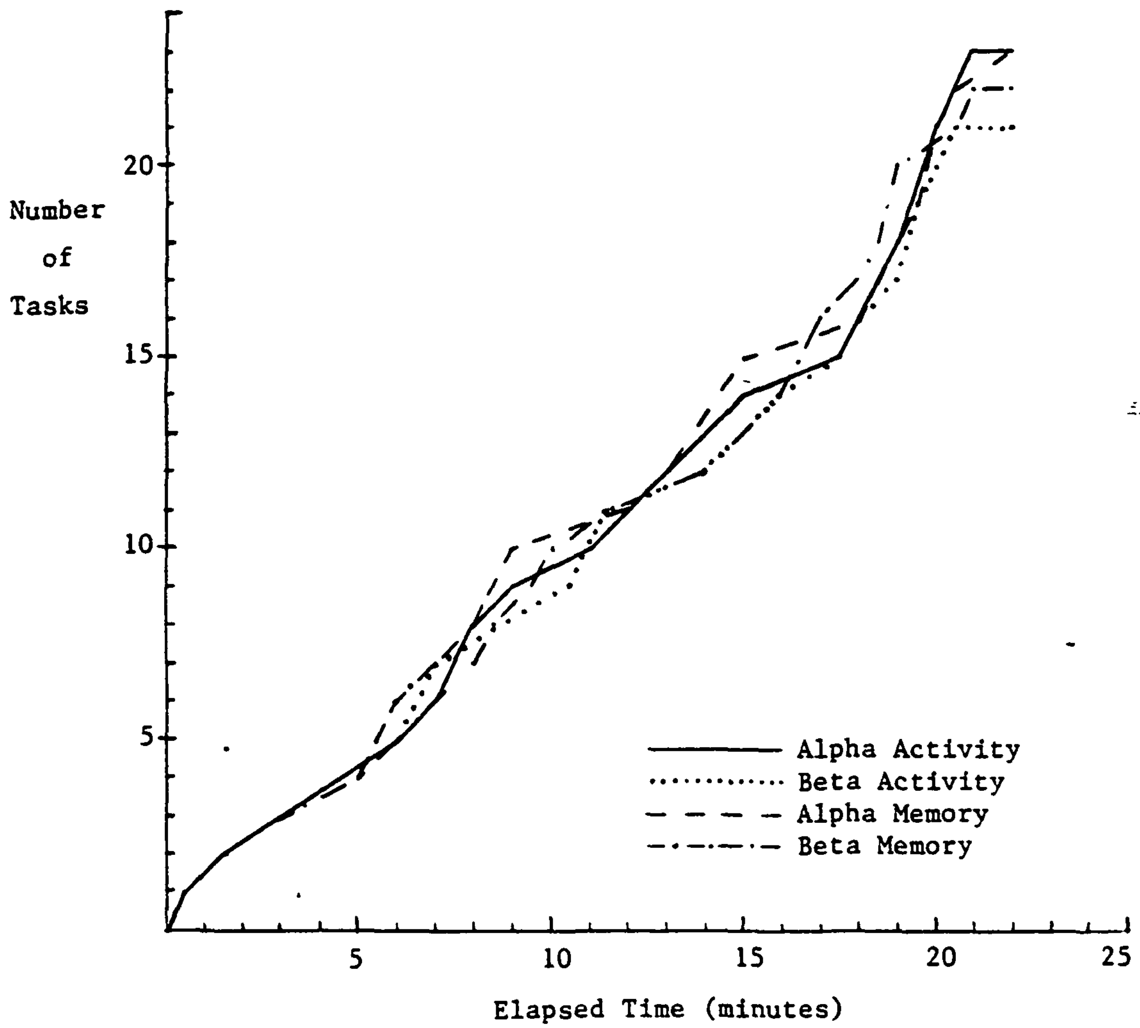


Figure 26: Accumulated Total Number of Memory Tasks

ORIGINAL PAGE IS
OF POOR QUALITY

The experiment you are participating in will provide information on pilot workload. The experiment consists of four "flights": two now and two another day. On each day you will fly two different ground tracks, terminating in an ILS approach. For each flight, the number of manual and mental tasks will be varied.

Your task is to fly as precisely as possible while following instructions to the best of your ability.

Ignore any ATC statements or instructions which appear on the display. All instructions and ATC statements will be handled verbally. However, when contacting a new "Controller", toggle off (away) the old radio and toggle on (toward) the new channel. Since all flights will be performed manually, you can ignore the two autopilot controls. In addition, the Trim and CWS switches are best left as set.

You will use 3 Navigation aids: VOR 1, VOR 2, and ILS 4. ILS 4 provides an ILS for Runway 36. Please note that the signal is only received within 10 miles of the runway. So, when on a dogleg to the ILS, hold heading until the Course Deviation Bar comes off the stops or the Glide Slope Indicator shows movement.

The "nominal" airspeed for these runs is 200 kts. Final approach will be flown at 150 kts. with Gear and Flaps down. Usually, a throttle position near center will maintain a stable airspeed.

You can expect the following level flight attitudes:

200 kts:	Clean	-2 deg
	Flaps	-5 deg
	Gear & Flaps	-2 deg
150 kts:	Clean	0 deg
	Flaps	+2 deg
	Gear & Flaps	+6 deg

During and after each run, you will be asked to make several subjective ratings. Thank you for your time and effort.

Figure 27: Pilot Instructions for the Preliminary Experiment

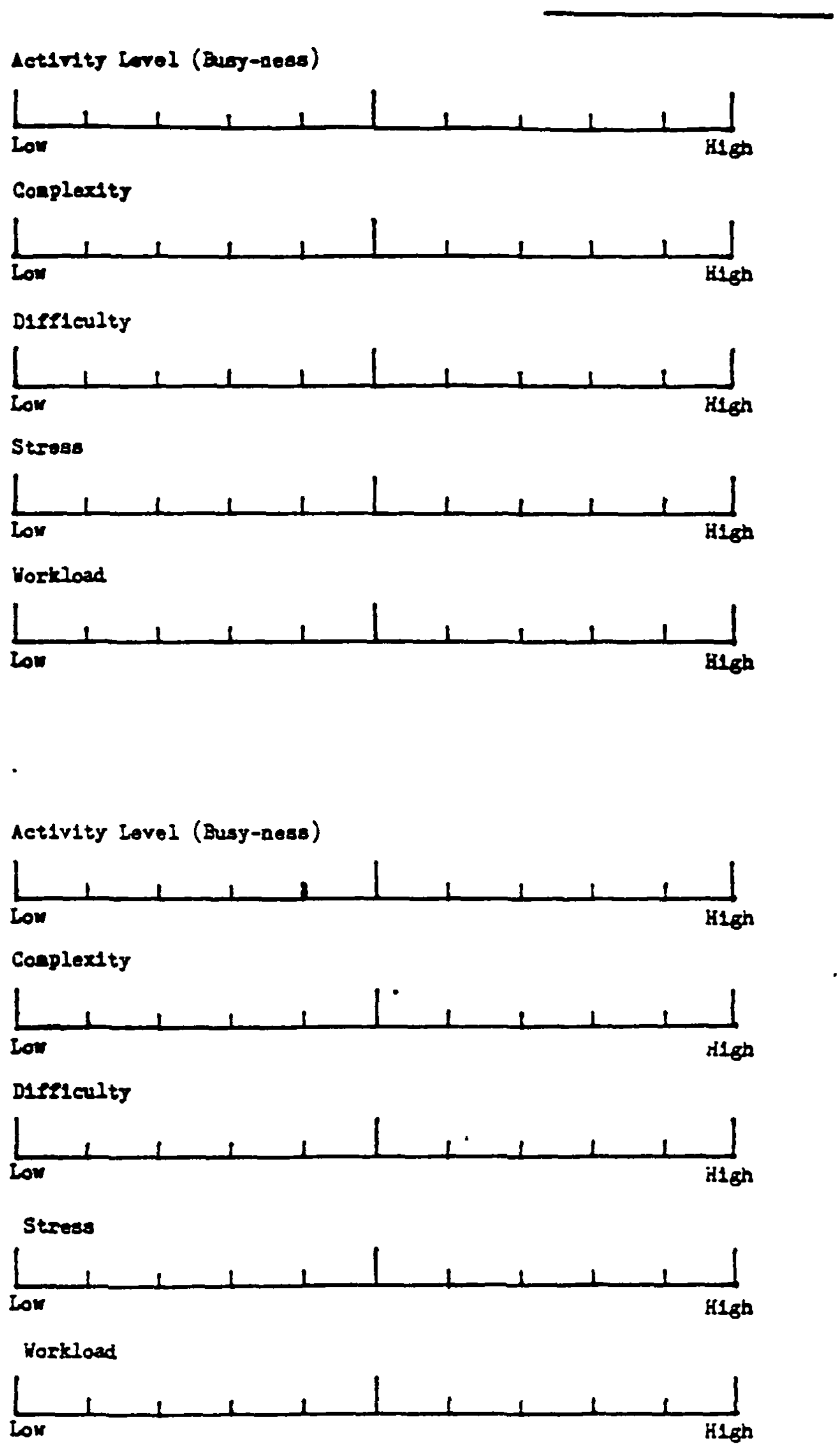


Figure 28: Subjective Rating Sheet for the Preliminary Experiment

ORIGINAL PAGE IS
OF POOR QUALITY

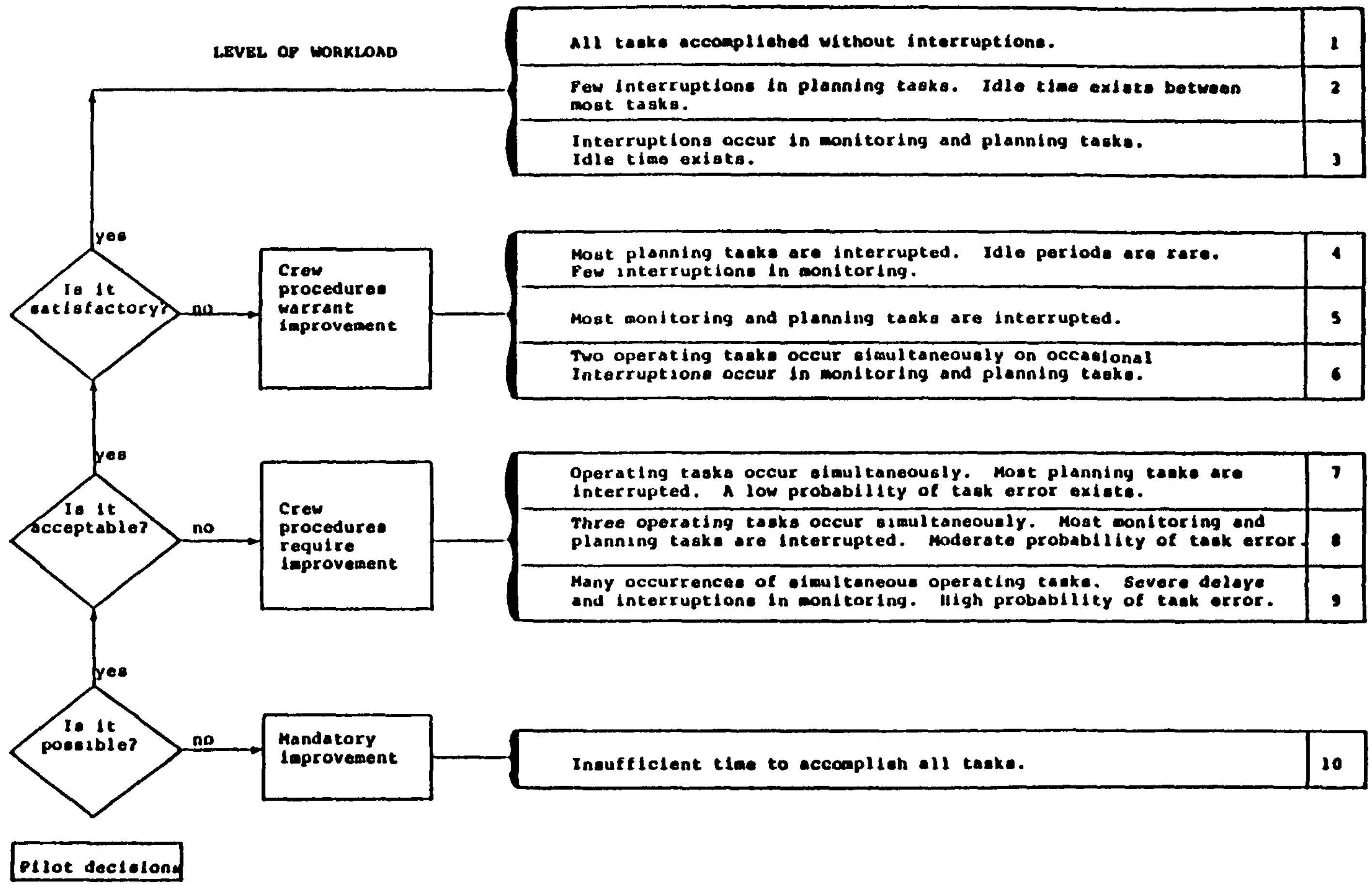


Figure 29: Reference example for rating subjective workload

activity: two runs. Volunteers were told that they would be asked to make ratings three times during each run and were to mark a "1" for their first rating, a "2" at their second rating, and a "3" at their last rating. In addition, they were asked to place a "1" for their overall rating at the end of each run. Thus, three times during each run the simulation was halted and the subjects rated Activity Level, Complexity, Difficulty, Stress, and Workload.

Figure 29 provided a reference for rating the Workload level. This modified Cooper-Harper system was adapted from earlier work by Sheridan and Simpson [24]. The validity and utility of this system was demonstrated by Katz [12].

The data runs were interrupted at 8 to 10 minutes and 18 to 20 minutes elapsed time. These two periods and run termination were used for ratings. After each run, the pilots were debriefed. They were asked for verbal or written comments concerning their ratings, performance, or actions.

3.4 DATA

Every 10 seconds, the computer stored aircraft x, y, and z positions. In addition, it stored every control box manipulation along with the magnitude and time of the event. This data provided ground track information. By comparing elapsed time with a time versus altitude profile, desired altitude was determined. Desired altitudes were then compared with the aircraft's actual altitudes to derive altitude error data. No altitude errors were computed during directed climbs and descents.

Any one of a multitude of reasons might cause the actual ground track to differ from the projected nominal ground track. For example, one pilot might lead a turn more than another, or use a slightly different course intercept heading. Therefore, ground track deviations were not computed. However, all ground tracks were plotted as a record of unusual activity, since major errors would manifest themselves.

Altitude data was chosen as an objective measure rather than airspeed data for several reasons. First, the altitude range was far greater. Altitudes ranged from sea level to over 5000 feet. Airspeeds ranged from 150 knots (kts) to 225 kts. Second, the range of potential altitude deviations was greater than airspeed deviations. Prior to flying the ILS, a pilot would need a deviation of at least 1500 feet to crash. However, once configured at 150 kts, only a 20 kt deviation (130 kts) would be required to stall the aircraft. Third, "anchoring" was available for all the desired altitudes, but not all of the airspeeds. That is, all the altitudes which the pilots were instructed to maintain resulted in the

altitude indicator being coincident with a mark on a dial. However, some airspeeds, such as 225 kts, resulted in placing the airspeed indicator half way between two marks. Thus, it was easier to be precise in interpreting present altitude than present airspeed.

This altitude error data was converted into Absolute Altitude Error data (feet) and Root-Mean-Square (RMS) Altitude Error data (feet). As mentioned earlier, the volunteer pilots were given no clues about which parameters would be used to measure performance.

Finally, previous work by Katz [12] and others had determined that no single subjective rating was an adequate mental workload measure. So, each pilot made five Subjective Workload Ratings at three points during each run. No ratings were made by independent observers since previous work had shown that pilots proved to be as reliable as observers in making these ratings. Activity Level, Complexity, Difficulty, Stress, and Workload were rated. Ratings were taken at three points rather than taking one overall rating to see if any segment was "point" loaded relative to the others.

A combination of subjective ratings and objective measures was used for several reasons. First, mental workload is generally agreed to be multi-dimensional in nature. Thus, multiple measures should provide a more complete picture of operator behavior and performance. Second, prior research attests to the importance and necessity of combining objective and subjective data to derive meaningful results. Hicks and Wierwille [8] stressed that both measures "...should be used in assessing...workload, particularly if it (the task) is of a psychomotor nature."

3.5 RESULTS

Table 2 lists the subjective ratings which the four pilots gave during their "Activity" flights. Ratings are given for all five subjective measures and each of the three rating periods. Alpha and beta route ratings are also shown. Table 3 shows the same data, but for the "Memory" flights. This information is summarized in Tables 4 and 5.

Student t-tests and F-tests were performed on the data. For the activity version scenarios and the memory version scenarios, there was no significant difference between the alpha or beta routes at a 95 percent confidence level for <u>any</u> of the five subjective categories. Either the design strategy was successful in minimizing differences between the two routes, or these measures were not sufficiently sensitive to demonstrate a difference.

For each of the four scenarios (for example, alpha/activity or beta/memory), there was no significant difference in the ratings for

Table 2: Activity version Subjective ratings

	Point	1		2		3	
	Route	a	B	a	B	a	B
	Subject						
Activity Level	A	5.0	3.0	3.0	5.0	3.0	5.0
	B	5.0	7.0	4.0	5.0	4.0	5.0
	C	7.2	6.2	6.4	4.6	7.5	5.4
	D	4.0	2.1	5.0	2.0	5.0	2.0
Complexity	A	3.0	1.6	2.0	5.0	2.6	5.0
	B	5.0	6.0	4.0	5.0	4.0	6.0
	C	6.4	6.3	6.7	5.2	7.3	6.7
	D	3.0	1.0	3.0	1.1	3.0	1.3
Difficulty	A	4.0	2.4	2.5	5.0	2.5	6.6
	B	5.0	6.0	4.0	6.0	4.0	6.0
	C	6.4	5.8	6.1	4.6	7.0	6.3
	D	6.0	1.1	6.0	2.0	7.0	1.2
Stress	A	5.0	1.5	3.0	4.0	2.0	5.5
	B	5.0	6.0	4.0	6.0	4.0	6.0
	C	7.3	6.2	6.3	5.3	8.3	5.6
	D	3.1	1.1	3.2	1.1	3.3	1.2
Workload	A	4.0	3.0	3.0	3.0	2.0	4.8
	B	6.0	8.0	5.0	7.0	5.0	8.0
	C	7.0	5.7	6.0	4.6	8.0	5.3
	D	3.0	1.0	5.0	1.1	6.0	2.1

Table 3: Memory version Subjective ratings

	Point	1		2		3	
	Route	a	B	a	B	a	B
	Subject						
Activity Level	A	1.0	2.0	4.0	3.0	5.0	2.2
	B	4.0	3.0	7.0	4.0	5.0	3.0
	C	3.7	4.2	4.5	3.4	5.0	6.4
	D	1.0	2.0	2.0	2.0	4.0	3.0
Complexity	A	1.0	1.0	5.0	1.5	3.0	1.3
	B	3.0	3.0	7.0	4.0	5.0	3.0
	C	4.3	2.8	5.0	3.3	5.6	6.4
	D	1.0	1.9	1.0	2.1	1.0	3.0
Difficulty	A	1.0	1.5	5.0	2.0	5.0	1.8
	B	3.0	3.0	7.0	4.0	5.0	3.0
	C	4.2	4.3	4.8	3.8	5.7	7.3
	D	1.0	1.9	2.0	2.1	4.0	3.0
Stress	A	1.0	1.0	5.0	1.5	5.0	1.3
	B	3.0	3.0	7.0	3.0	5.0	3.0
	C	3.4	5.6	4.1	4.8	3.8	7.3
	D	1.0	1.8	1.0	2.0	3.0	3.0
Workload	A	1.0	1.5	4.0	2.0	2.5	2.0
	B	2.0	2.0	8.0	3.0	7.0	2.0
	C	4.4	5.4	4.8	5.6	5.8	6.7
	D	1.0	2.0	2.0	3.0	3.0	3.0

Tables 4 and 5: Overall Subjective ratings

ACTIVITY SCENARIOS

	mean			standard deviation		
	a	B	a&B	a	B	a&B
Activity Level	4.9	4.3	4.6	1.4	1.6	1.5
Complexity	4.1	4.1	4.1	1.7	2.1	1.9
Difficulty	5.0	4.3	4.6	1.5	2.1	1.9
Stress	4.5	4.1	4.3	1.8	2.1	1.9
Workload	5.0	4.4	4.7	1.6	2.3	2.0

MEMORY SCENARIOS

	mean			standard deviation		
	a	B	a&B	a	B	a&B
Activity Level	3.8	3.1	3.5	1.6	1.2	1.5
Complexity	3.4	2.7	3.1	2.0	1.3	1.7
Difficulty	3.9	3.1	3.5	1.7	1.5	1.7
Stress	3.5	3.1	3.3	1.8	1.8	1.8
Workload	3.7	3.1	3.4	2.1	1.6	1.9

segments 1, 2, or 3 at a 90 percent confidence level. This implied that for these scenarios, the overall workload varied little from phase to phase.

Mean Subjective Rating data is plotted in Figure 30 for all five categories. Since there was little difference between the alpha and beta routes, I have only plotted the overall mean ratings for the activity and memory versions. The mean memory ratings are shown as circles, and the mean activity ratings are shown as triangles.

Student t-tests showed a statistically significant difference between the two versions for the Complexity and Stress ratings at a 90 percent confidence level. The difference was significant at a 95 percent level for Activity Level, Difficulty, and Workload.

The lower confidence level for the Complexity rating may be due to the fact that all runs were performed manually. That is, the autopilot was not used. Thus, "complexity" changed little. The lower confidence level for Stress may be due to the relatively low workload level for these experiments. These experiments attempted to simulate a normal air traffic environment. The relatively low ratings are, therefore, consistent with Katz's [10] findings for a similar environment.

As Figure 30 shows, the activity version ratings were consistently higher (harder, more difficult) than the memory version ratings. This was somewhat surprising since the average <u>total</u> (activity plus mental) WU's were greater for the memory version than the activity version. (218.5 WU's versus 187.0 WU's: 116.8 percent)

Since other results had lent credibility to the use of this "workload unit" technique, several explanations are possible. First, the 17 percent difference in WU's between the two versions may not be significant at these low to moderate workload levels. Second, because subjects were "busier", doing a greater number of relatively simple tasks, they may have equated simple busyness with greater workload. (This premise, however, would contradict Katz's [12] results in a similar experiment.) Third, the nature of physical and mental workload may be different enough to invalidate comparing workload levels by simply adding the two types of WU's. Fourth, the mental "workload unit" model may be faulty. There may be relatively heavy mental workloads initially and at information retrieval, with little or very low mental workload in between.

Tables 6 and 7 give the altitude error data for the activity and memory versions. The table lists mean errors, the standard deviation, and rms errors for each subject and both alpha and beta routes. The data is also averaged across all the pilots and across both routes. This information is shown for the overall mean and rms altitude deviations in Figure 31.

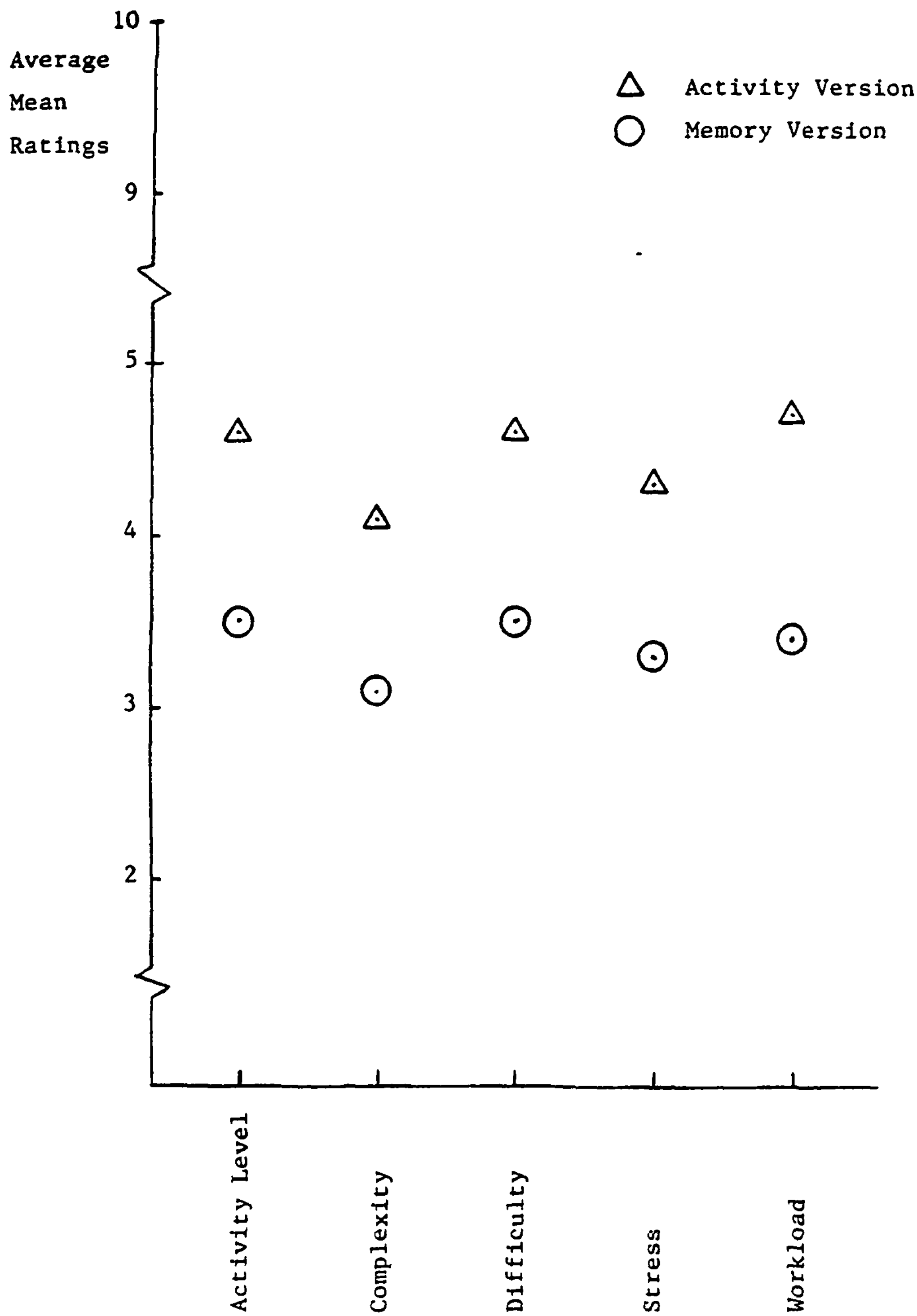


Figure 30: Average mean subjective ratings by category

Tables 6 and 7: Altitude Error Data (feet)

ACTIVITY VERSION

		Pilots				avrg.	overall	
		A	B	C	D			
a	mean	59.0	77.3	95.7	69.4	74.3		
	std dev	36.1	71.4	124.7	59.3	79.9		
	rms	69.1	105.1	157.2	91.3	109.2	74.4	mean
B	mean	60.6	95.0	59.1	99.0	74.6	76.1	std dev
	std dev	53.3	57.6	60.6	105.8	71.1	106.5	rms
	rms	80.7	111.0	84.7	144.9	103.0		

MEMORY VERSION

		Pilots				avrg.	overall	
		A	B	C	D			
a	mean	81.7	163.0	72.4	121.6	103.0		
	std dev	49.7	134.8	57.4	94.3	90.3		
	rms	95.6	211.5	92.4	153.9	137.0	93.5	mean
B	mean	93.2	122.3	74.3	53.4	84.1	81.0	std dev
	std dev	70.8	80.6	61.6	45.6	69.4	123.8	rms
	rms	117.0	146.4	96.5	70.2	109.1		

Student t-test analysis of these errors gives the following results: (1) mean absolute altitude errors were different for the activity and memory versions at an 80 percent confidence level; (2) rms altitude errors were different at a 70 percent confidence level.

The relative weakness in distinguishing these two versions with objective data may be due to the fact that there was no "baseline" version for comparison. Both versions were designed to be difficult, but difficult in different ways. Furthermore, both versions were rated only moderately difficult. Experiments at a higher level of difficulty may increase the sensitivity of this measure.

As Figure 31 illustrates, both types of altitude errors were greater for the memory versions than the activity versions. This is surprising since the activity version had a far more demanding altitude profile.

The greater altitude errors for the memory version are probably not due to pilot boredom. No individual run lasted more than 30 minutes, and runs were broken by several "freezes" for subjective ratings. Also, the pilots knew that their performance was being measured, increasing interest. Finally, the memory version had few "quiet" periods longer than several minutes.

Two other, more promising, explanations relate to interest or attention. In the activity version, subjects were repeatedly asked to change airspeed, altitude, and heading. Thus, they probably channelled more effort and attention to these tasks, resulting in smaller deviations. This would also help explain the slightly higher subjective ratings for this version.

Alternatively, another type of prioritizing may have occurred. Given a lower task workload in the memory version, the subjects may have shifted aircraft control to a lower priority. This would produce a certain level of complacency about altitude, while subjects paid additional attention to memory items.

Table 8 provides data on long-term memory errors for all four scenarios. However, this chart further differentiates among long-term memory tasks: Positional tasks and Non-Positional tasks. A "Positional" task pertains to something requiring changing the aircraft's state: for example, "Descend and maintain 3000 at Point Delta". An example of a "Non-Positional" task is, "Report at Point Delta".

Although it is difficult to generalize because of the small total number of tasks, the percentage of forgotten "Positional" tasks was similar for all four scenarios and the percentage of forgotten "Non-Positional" tasks was also similar for all four scenarios. However, on average, only 12.5 percent of "Positional"

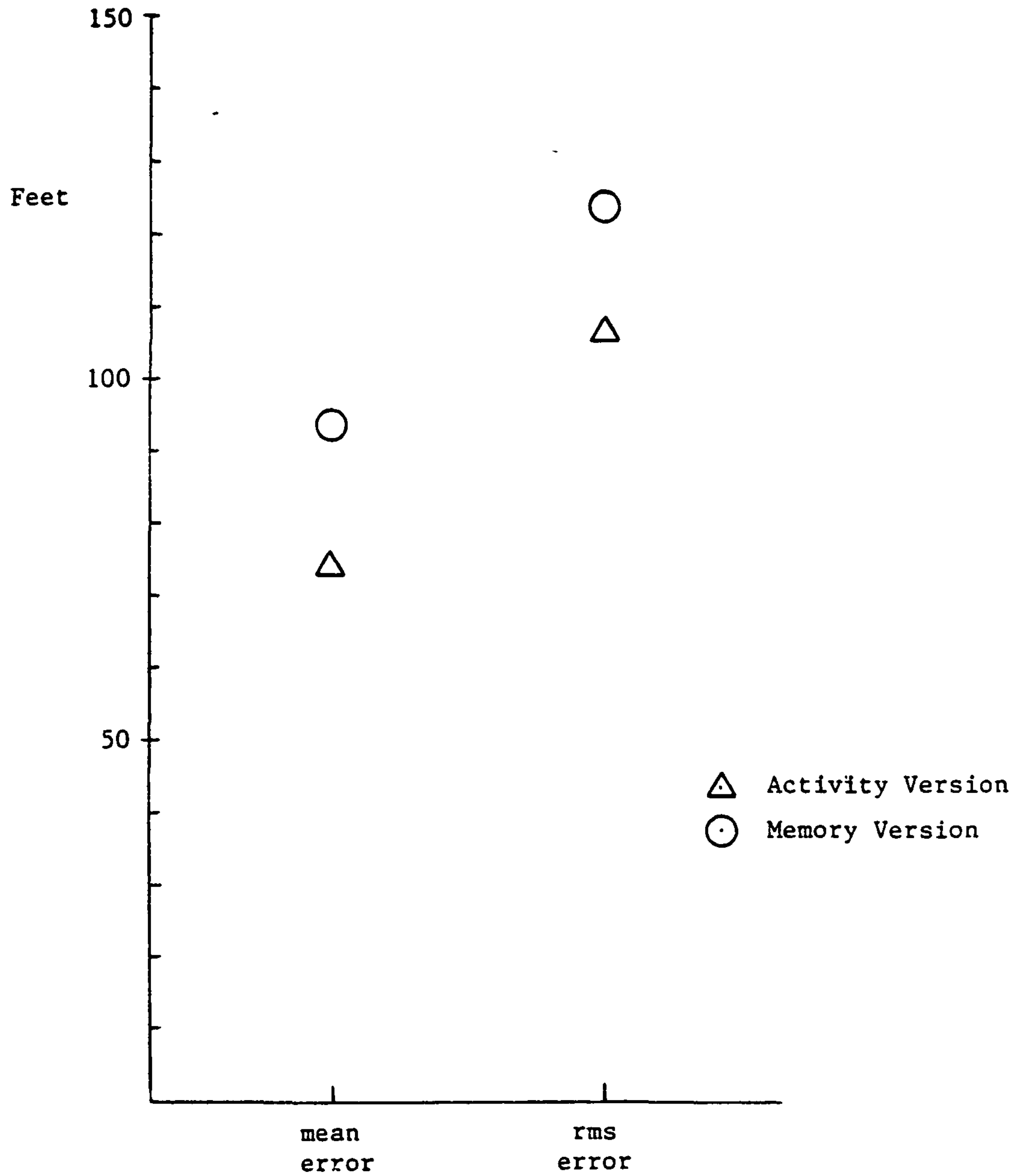


Figure 31: Average altitude deviations for each version

tasks were missed, while 40.6 percent of "Non-Positional" tasks were missed.

Table 8: Long-term Memory Errors:
Positional and Non-Positional

Scenario	Activity a	Activity B	Memory a	Memory B
Positional Tasks	0	0	8	8
Number of Errors	0	0	1	1
Error Percentage	0	0	12.5	12.5
Non-Positional Tasks	4	4	12	12
Number of Errors	2	2	5	4
Error Percentage	50.0	50.0	41.7	33.3

Thus, it appears that the pilots were deliberately prioritizing memory items by type. Requirements relating to aircraft state received higher priority than ARTCC requests, for example. These results are consistent with a study by Loftus, Dark, and Williams [15]. They found that "place" information was well remembered while "frequency" information was remembered relatively poorly.

3.6 MAIN FINDINGS FROM PRELIMINARY EXPERIMENT

1. The Workload Unit (WU) technique appears to work satisfactorily for quantifying similar types of workload. It works less well for comparing dissimilar (i.e. mental and physical) workloads.

2. At low to moderate workload levels, pilots reported a higher workload when given physical tasks than when given memory tasks.

3. At a low to moderate workload level, pilots reported higher subjective workload ratings for scenarios where altitude deviations decreased, possibly due to greater attention or interest.

4. Greater memory workload appears to interfere with activity

performance.

5. At low to moderate workloads, subjective ratings were more sensitive to scenario differences than objective, primary-task measures. (This result is similar to that reported by Williges and Wierwille. [31])

6. Pilots systematically weighted aircraft state requirements higher than other requirements.

3.7 FOLLOW-ON EXPERIMENT IDEAS

These findings led to a number of conclusions and ideas relating to the next series of experiments. The new experiments would attempt to clarify and expand on the preceding findings.

The Workload Unit technique for quantifying workload had demonstrated its validity in certain limited applications. Thus, it would be used again.

A "Baseline" scenario would be added to provide a nominal scenario for comparison. It would be a low workload scenario, representative of routine terminal approach activity at a non-congested airfield.

The non-nominal scenarios would be designed to be much more difficult than the scenarios for these first experiments. Workload ratings on a 10 point scale exceeded 5 only 38 percent of the time, 6 only 20 percent of the time, and 7 only 7 percent of the time. A rating of 8 was exceeded only once in 120 ratings, and 9 never. Thus, there seemed to be a good deal of workload capacity remaining in the pilot volunteers.

It was decided that greatly increasing the workload level might increase the ability of objective measures to distinguish between a high physical tasking scenario and a high mental tasking scenario. It could also shed additional light on the multi-dimensional character of mental workload by producing significant differences among the various subjective ratings.

The differences in pilot performance for remembering "positional" versus "non-positional" memory tasks was striking. The next series of experiments could further examine this issue by increasing the mental workload and expanding the number of memory tasks.

Although some of these pilots were "fighter-types" and some were "heavy-types", all had a great deal of high-performance jet aircraft experience. The next series of experiments would expand

the experience base as much as possible. This could broaden the range of subjective workloads and produce more examples of "saturated" pilots.

Finally, the next series of experiments might examine the time sensitivity of mental tasks. That is, how much more likely are pilots to forget tasks far in the future than tasks which are closer at hand?

Chapter 4

PRIMARY EXPERIMENT: DESIGN

4.1 SUBJECTS

The results of the first set of experiments led me to recruit pilots with a wide range of experience. Katz [12] and others had found that pilots with lower levels of experience tended to rate workload higher than did more experienced pilots. So, I hoped that the less experienced pilots might be more easily "saturated", providing important data on how mental workload affected performance under these extreme conditions. Nevertheless, due to the nature of this series of experiments, even the less experienced pilots needed to be very proficient in instrument flying procedures.

Initially, approximately 30 pilots volunteered to participate. They were brought in to fly the simulator for at least one and sometimes two, 2-hour evaluation sessions. Although I had hoped to use at least a dozen pilots of varied background, the list of 30 was soon reduced to 10. Few of the original 30 had logged any high performance aircraft time, and the simulation's higher airspeeds and unfamiliar instrumentation disqualified most of these pilots.

Three of my ten finalists were eventually forced to withdraw from the experiment before finishing it. A lack of time and other commitments made it impossible for them to devote the number of hours or days necessary to practice, qualify on the simulator, and take part in all the data runs.

I ended up with seven pilots. All were good pilots, and there was a good mix of experience. Three were Air Force pilots with a great deal of flight time. Two pilots were Certified Flight Instructors with instrument ratings. The four civilian pilots ranged in experience from 300 total hours to 3000 total hours and had between 50 and 250 hours of instrument time.

The following is a breakdown of their experience:

A:	Light Aircraft:	....	300 Hours
	Total:	....	300
B:	Light Aircraft:	....	320
	Total:	....	320

C: Light Aircraft:	....	1300 Hours
Total:	....	1300
D: Sailplane:	....	1000
Light Aircraft:	....	2000
Total:	....	3000
E: Fighter-Type:	....	3200
Jet:	....	2750
Total:	....	3200
F: Light Aircraft:	....	500
Fighter-Type:	....	1200
Heavy Aircraft:	....	700
Jet:	....	1900
Total:	....	2400
G: Light Aircraft:	....	250
Fighter-Type:	....	500
Heavy Aircraft:	....	1750
Jet:	....	2250
Total:	....	2500

4.2 EXPERIMENTAL DESIGN

Only one route was used for this series of experiments. However, there were four scenarios once again. The four scenarios differed enough that it was felt only one route was necessary. Figure 32 illustrates the basic route.

The four scenarios were labeled Baseline, Activity, Planning, and Combined. The Baseline scenario was the easiest. It simulated a "normal" flight and the pilots were encouraged to use the autopilot to keep workload at a minimum. There were no directed deviations from the basic course, and airspeed and altitude changes were rare. Also, there were very few memory or planning tasks assigned.

A data session consisted of a Baseline run followed by one of

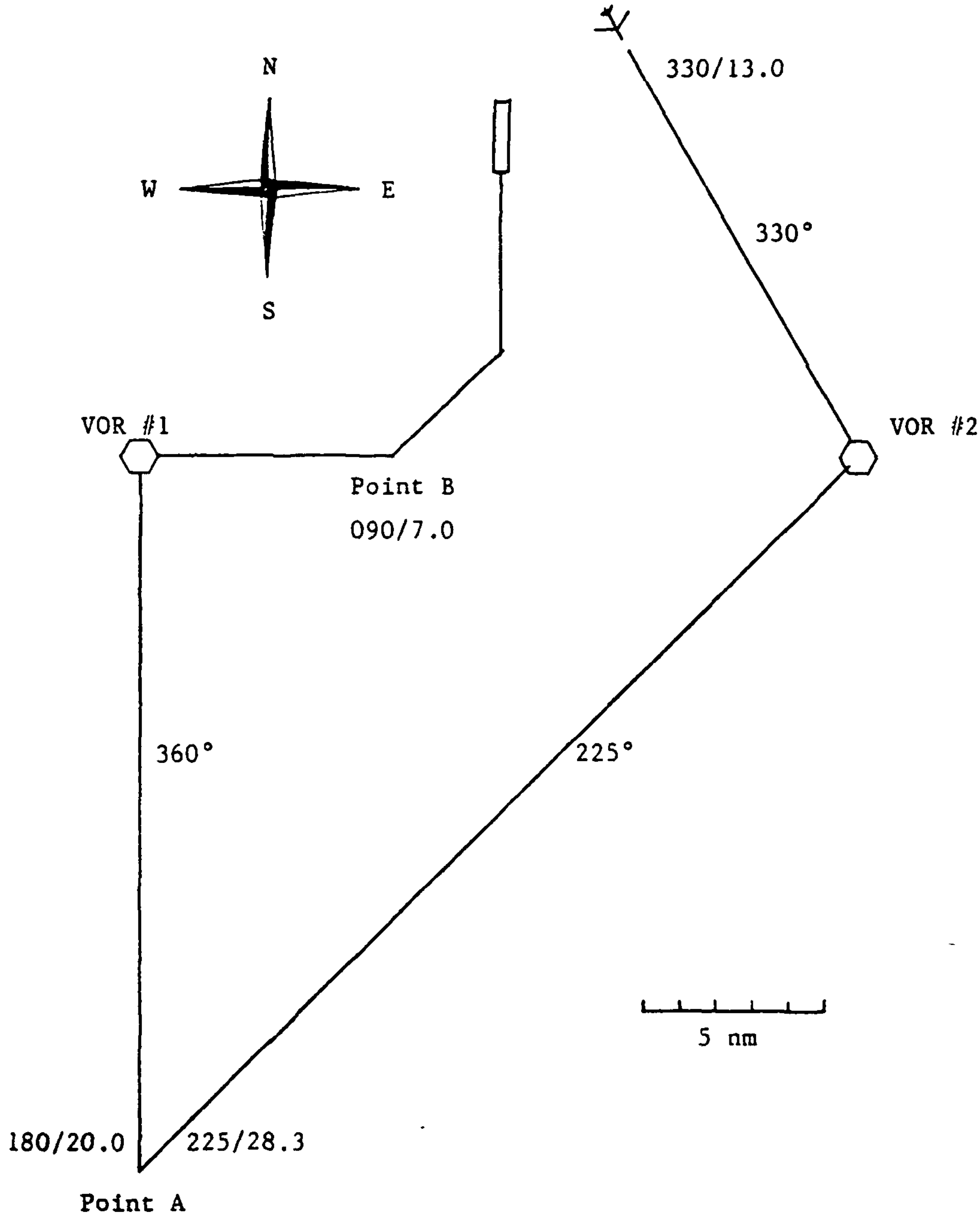


Figure 32: Navigation Chart for the Primary Experiment

the other scenarios. The Baseline scenario was used as a warm-up data run and as a calibration run. Each second run's data was compared to that session's Baseline run. Baseline performance and ratings for different sessions could then be compared to adjust the data for variations due to day-to-day differences such as fatigue, stress, emotional state, et cetera.

The Activity scenario was very similar to the Activity scenario of the preliminary experiment. It was loaded with a variety of manual-control tasks, but contained a planning task load similar to the Baseline scenario. The pilots flew this scenario without using the autopilot. It did differ significantly from the activity scenario of the preliminary experiments in that its activity workload (as measured in WU's per minute) was 40 percent greater while its memory workload was 50 per cent lower.

The Planning scenario was very different from the Activity scenario. It was almost identical in manual activity to the Baseline scenario, (and thus, had a low activity level) but instead of being directed to perform actions immediately, the pilots were directed to perform these actions at a certain time in the future. These instructions often involved overlapping time periods, and the requests were not ordered chronologically. Therefore, the pilots had to sort out the instructions and "plan".

For example, prior to 2:00 minutes the pilot might be told to descend 1000 feet at 5:00 minutes, then told to turn to 300 degrees heading at 13:30 minutes, then to slow to 190 knots at 8:00 minutes.

In terms of memory WU's, the Planning scenario was 80 percent more difficult than the Memory scenario of the preliminary experiments. Conversely, its activity WU rate was only one-third that of the preliminary runs. This scenario was flown on autopilot to keep the manual-control workload low.

The Combined scenario was designed to be the most difficult of all. It combined the manual activity of the Activity scenario with the planning requirements of the Planning scenario. This was an effort to saturate the pilots and see if any performance measure deteriorated sharply. The pilots were allowed to use the autopilot for help, but the pace of this scenario usually limited its use to making turns and holding headings.

Table 9 lists the order in which each pilot flew each of the non-Baseline scenarios. Different pilots flew the various scenarios in different orders. However, as mentioned earlier, they all began each session's data runs with a Baseline run. The other three scenarios were not truly order randomized, but they were mixed. No pilot flew the Combined scenario in the first session. It was so unusually difficult, it was felt that starting with this scenario might create an impossible workload for any pilot flying it first.

Therefore, all subjects flew either the Activity or Planning scenario in their first session. Then, the Combined scenario might be flown in either the second or third session.

Table 9: Session number in which pilots flew each scenario

SCENARIO	PILOT						
	A	B	C	D	E	F	G
Activity	1	2	3	3	1	2	1
Planning	2	1	1	1	3	1	2
Combined	3	3	2	2	2	3	3

A Navigation Chart (Figure 32) and a note pad were provided for each pilot's use. Also, special placards were displayed beneath the instrument display to give configuration/airspeed data and help the pilots with the various lateral and longitudinal autopilot modes.

Ground tracks, altitude profiles, and airspeed profiles provided in Figures 33 through 37, clearly illustrate some of the differences and similarities of the various scenarios. Those three items were nearly identical for the Baseline and Planning scenarios, and for the Activity and Combined scenarios. Figure 33 shows the ground track for the Baseline and Planning scenarios while Figure 34 shows the ground track for the Activity and Combined scenarios. Note the number of heading changes for the Activity/Combined scenarios. In the Activity and Combined scenarios the subjects were given new headings, altitudes, and airspeeds each 2 minutes for the first 5 minutes, each minute for the next 10 minutes, and each 30 seconds for the final 10 minutes. At several points, pilots were given instructions to contact ARTCC rather than perform some task.

Figure 35 is an airspeed versus time plot for the Planning and Activity scenarios. There are 31 airspeed changes for the Activity and Combined scenarios and 3 for the Baseline and Planning scenarios.

Figure 36 plots aircraft heading versus time. The Activity and Combined scenarios have 27 heading changes to 5 for the Baseline and Planning scenarios.

Finally, Figure 37 shows altitude versus time. The Activity and Combined scenarios have 21 directed altitude changes to 5 for

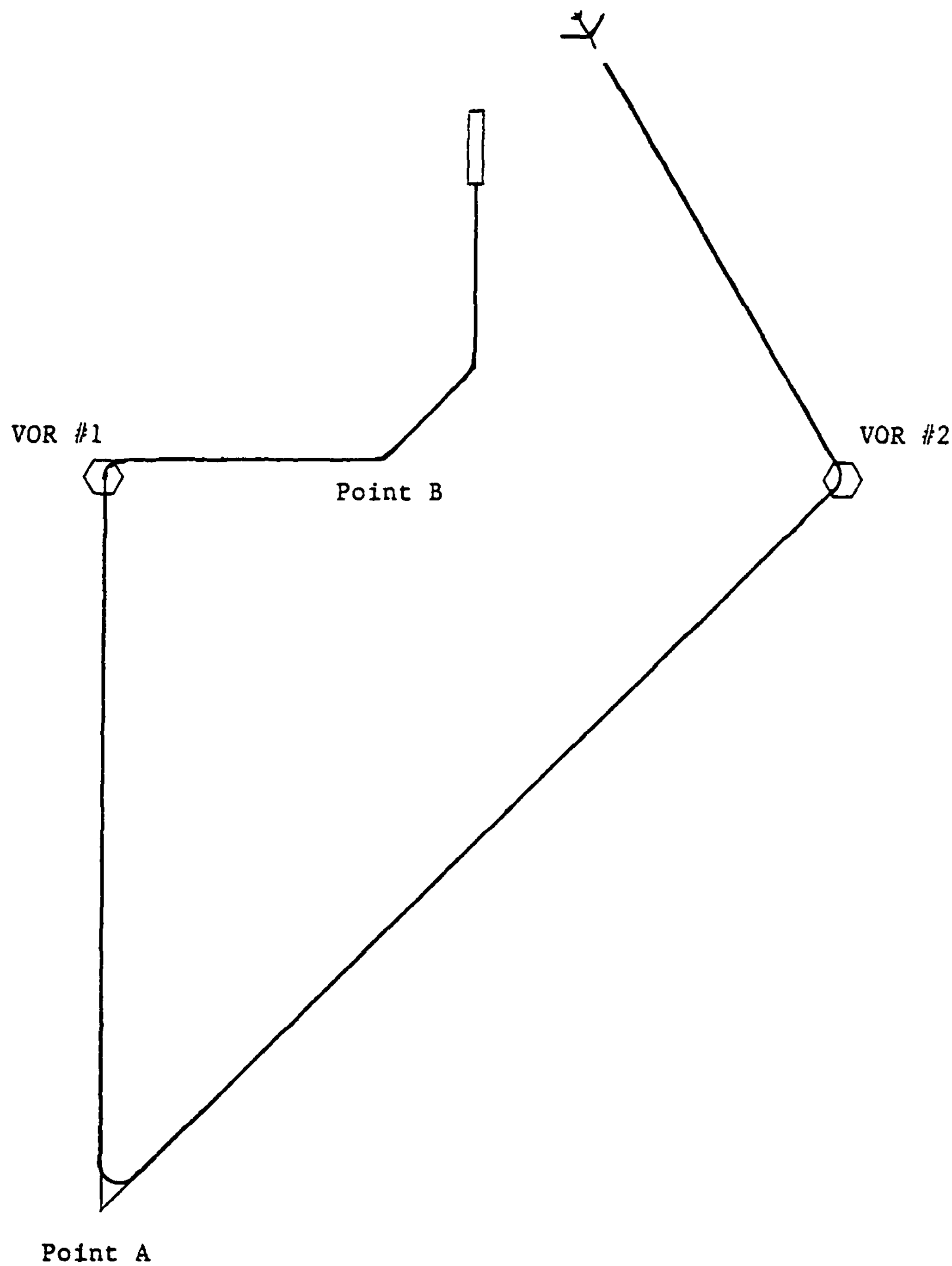


Figure 33: Nominal ground track for the Baseline and Planning scenarios

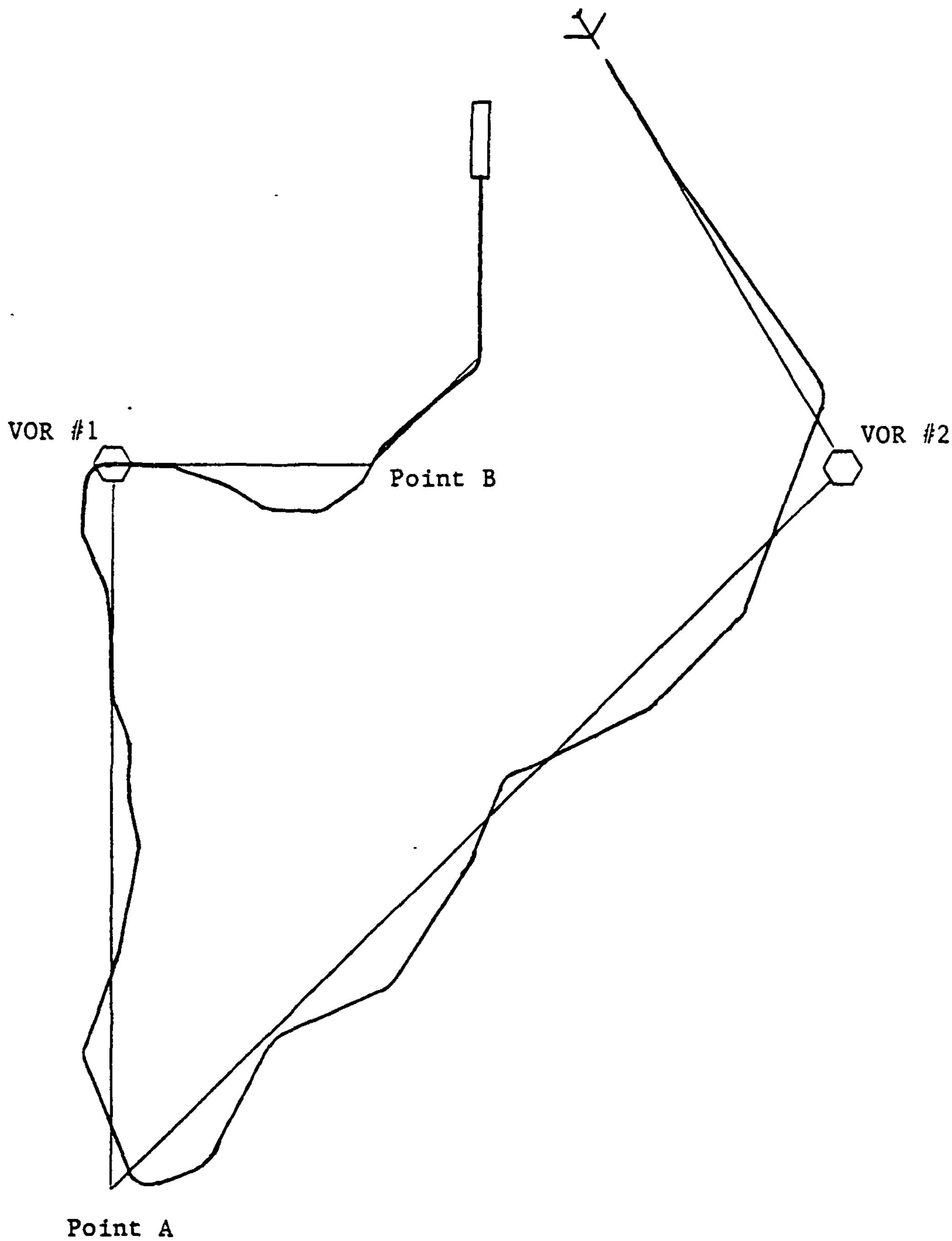


Figure 34: Nominal ground track for the Activity and Combined scenarios

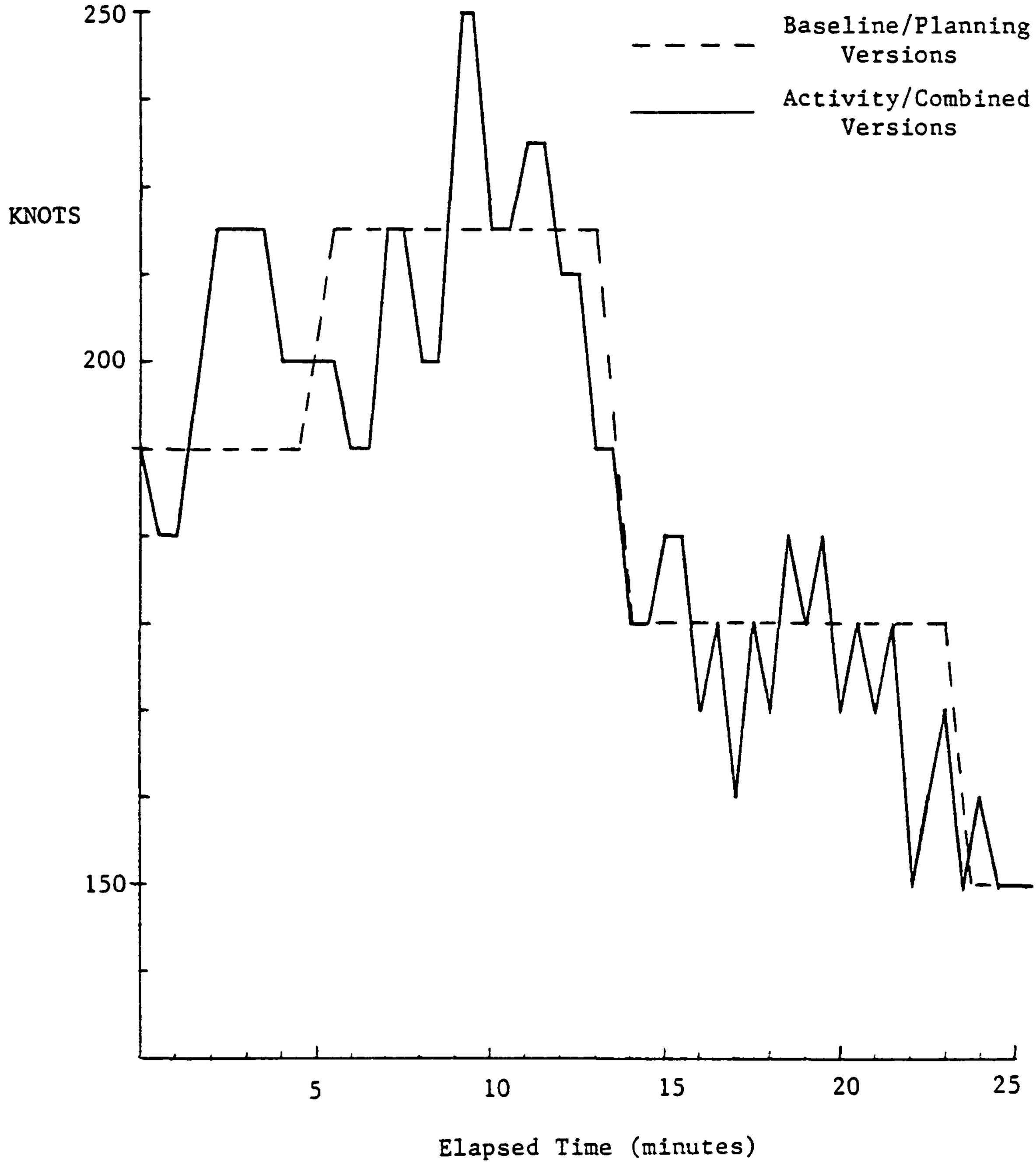


Figure 35: Planned airspeed versus elapsed time

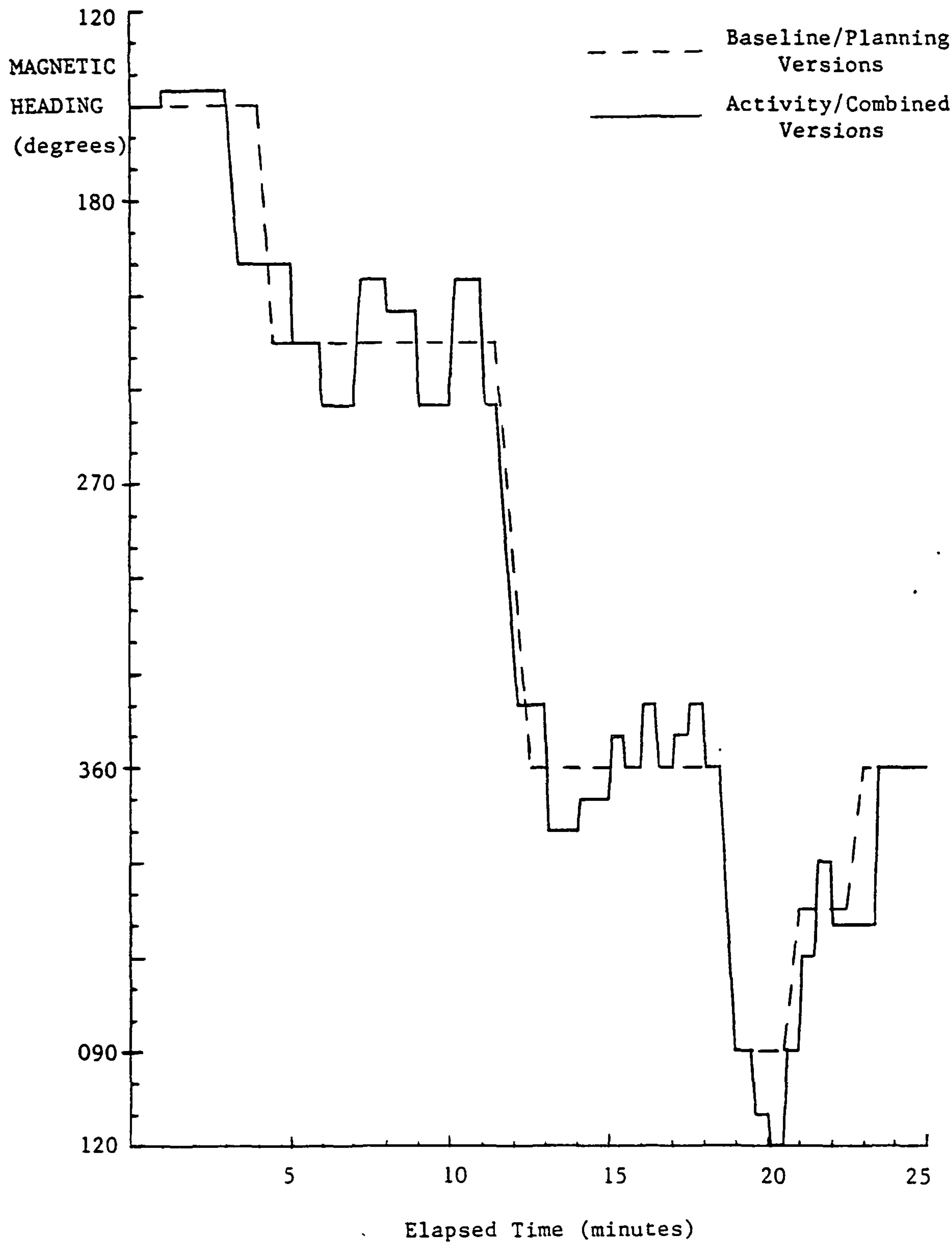


Figure 36: Planned magnetic heading versus elapsed time

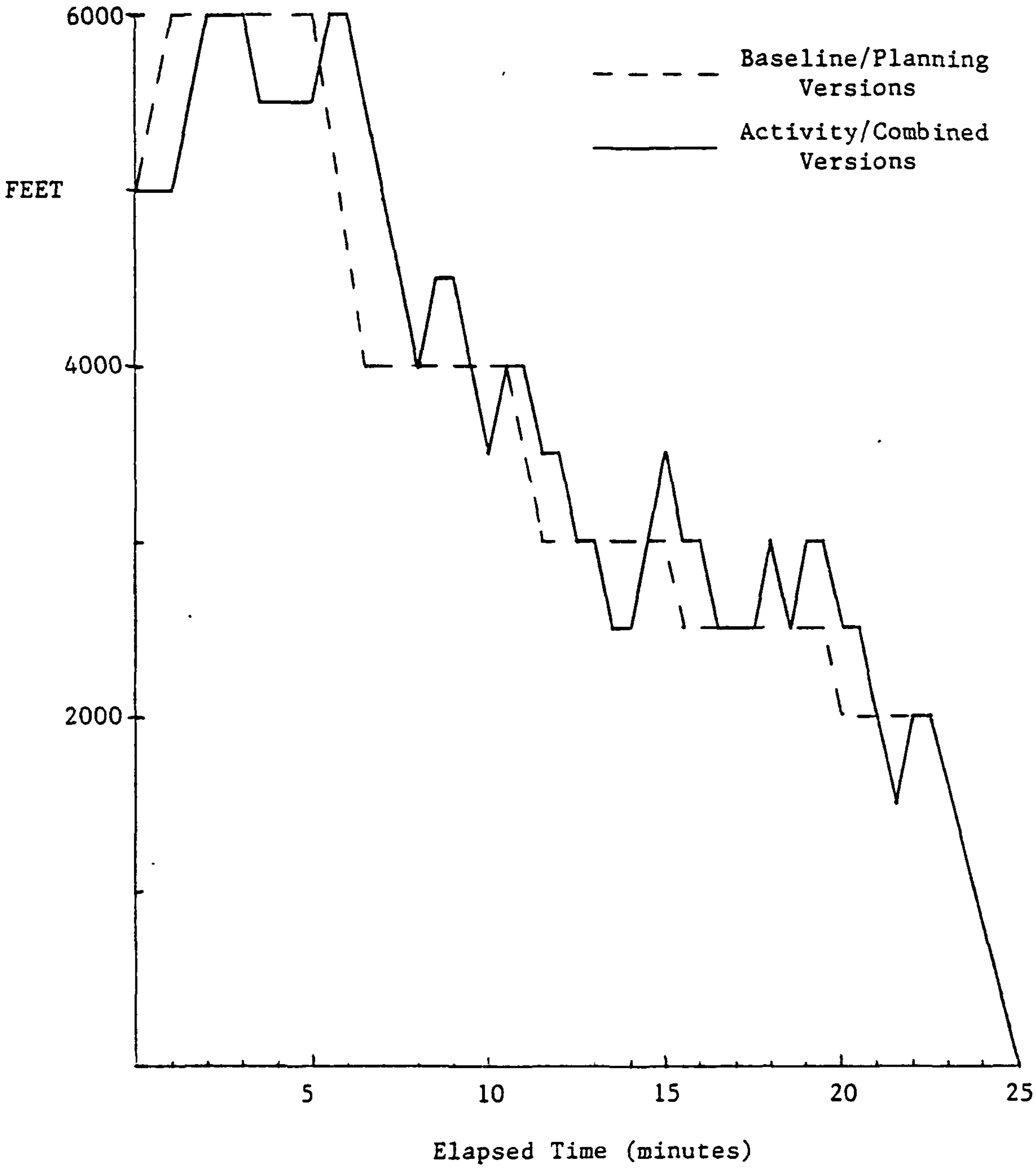


Figure 37: Planned altitude versus elapsed time

the Baseline and Planning scenarios.

Just as in the first set of experiments, workload units (WU's) were computed for the manual and mental tasks of the four scenarios. The procedure for calculating these relative workloads was the same as before. For a full description of the technique, see section 3.2.

Figures 38 and 39 are examples of the planning task and workload unit plots for each scenario. Figure 38 is for the Activity scenario and Figure 39 is for the Planning scenario. Each figure presents a variety of activities plotted against elapsed time. At the top, each square block represents carrying out one assigned planning task. Next is a plot of planning WU's. This is followed by a diagram showing the number and duration of short-term, medium-term, and long-term planning tasks. The bottom plot shows activity WU's.

I arbitrarily defined a short-term planning task as lasting from 0 to 4 minutes, a medium-term task lasting from 4 to 12 minutes, and a long-term task lasting over 12 minutes. The average short-term task was 2.6 minutes long, the average medium task was 7.2 minutes, and the average long-term task was 16.6 minutes.

Table 10 summarizes the information for all four scenarios. Note that the Planning and Combined scenarios have about 5 times as many planning WU's as the Baseline and Activity scenarios. Also, the Activity and Combined scenarios have roughly 5 times as many activity WU's as the Baseline and Planning scenarios. Finally, the Planning and Combined scenarios have almost 6 times as many planning tasks as the Baseline and Activity scenarios.

In recognition of Miller's [17] findings about human limits on immediate memory, the number of simultaneous planning tasks never exceeded 9. The Planning and Combined scenarios had an intense level of simultaneous planning tasks. However, the mean number of simultaneous planning tasks was only 5.0, with a standard deviation of 1.8.

Figures 40 and 41 portray some of this workload data graphically. Figure 40 is a plot of the accumulated number of activity WU's as a function of time. Figure 41 is a plot of the accumulated number of planning WU's as a function of time. Note not only the difference between dissimilar scenarios, but also the similar workload rate for similar scenarios.

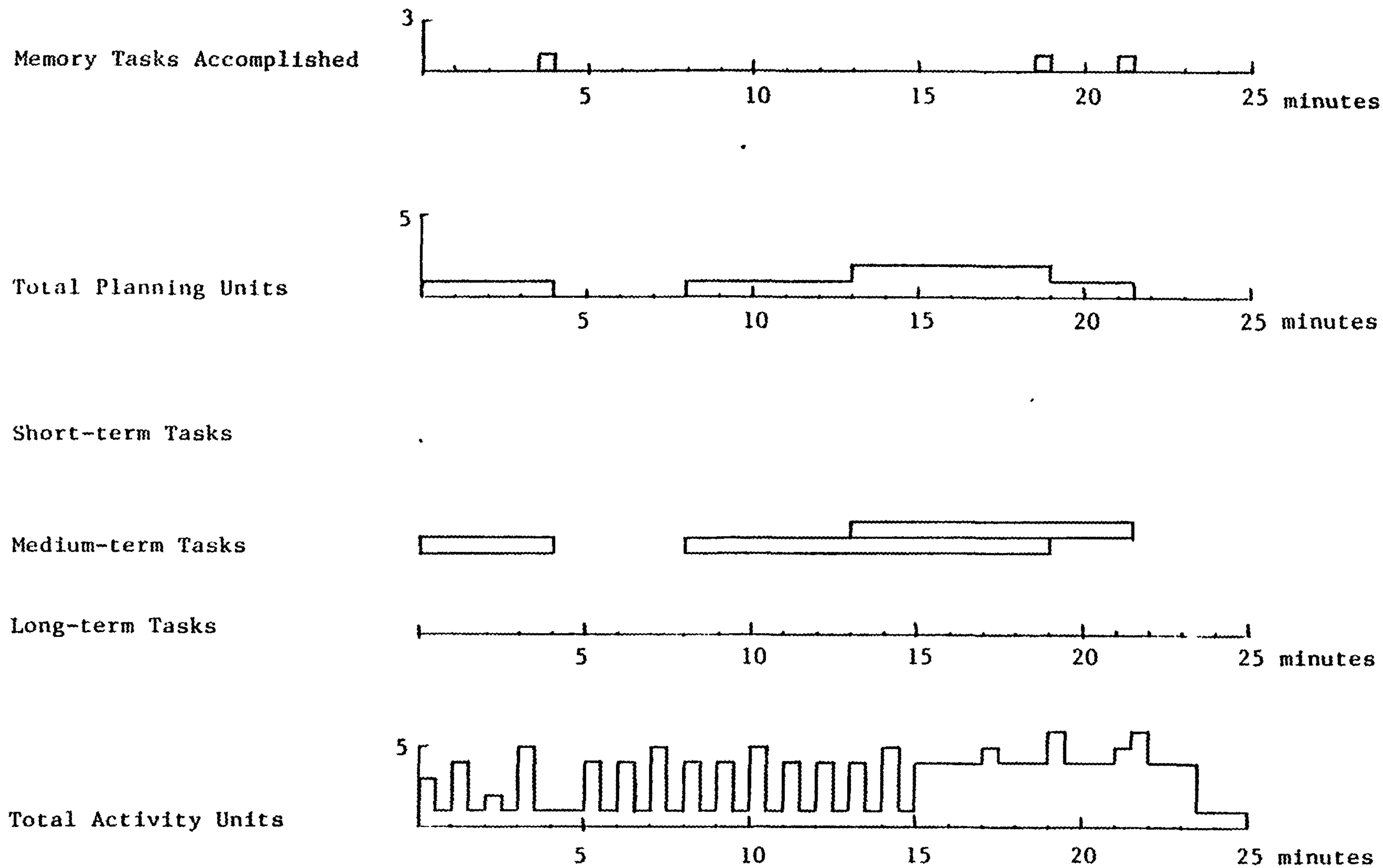


Figure 38: Activity scenario: Tasks and Workload Units versus elapsed time

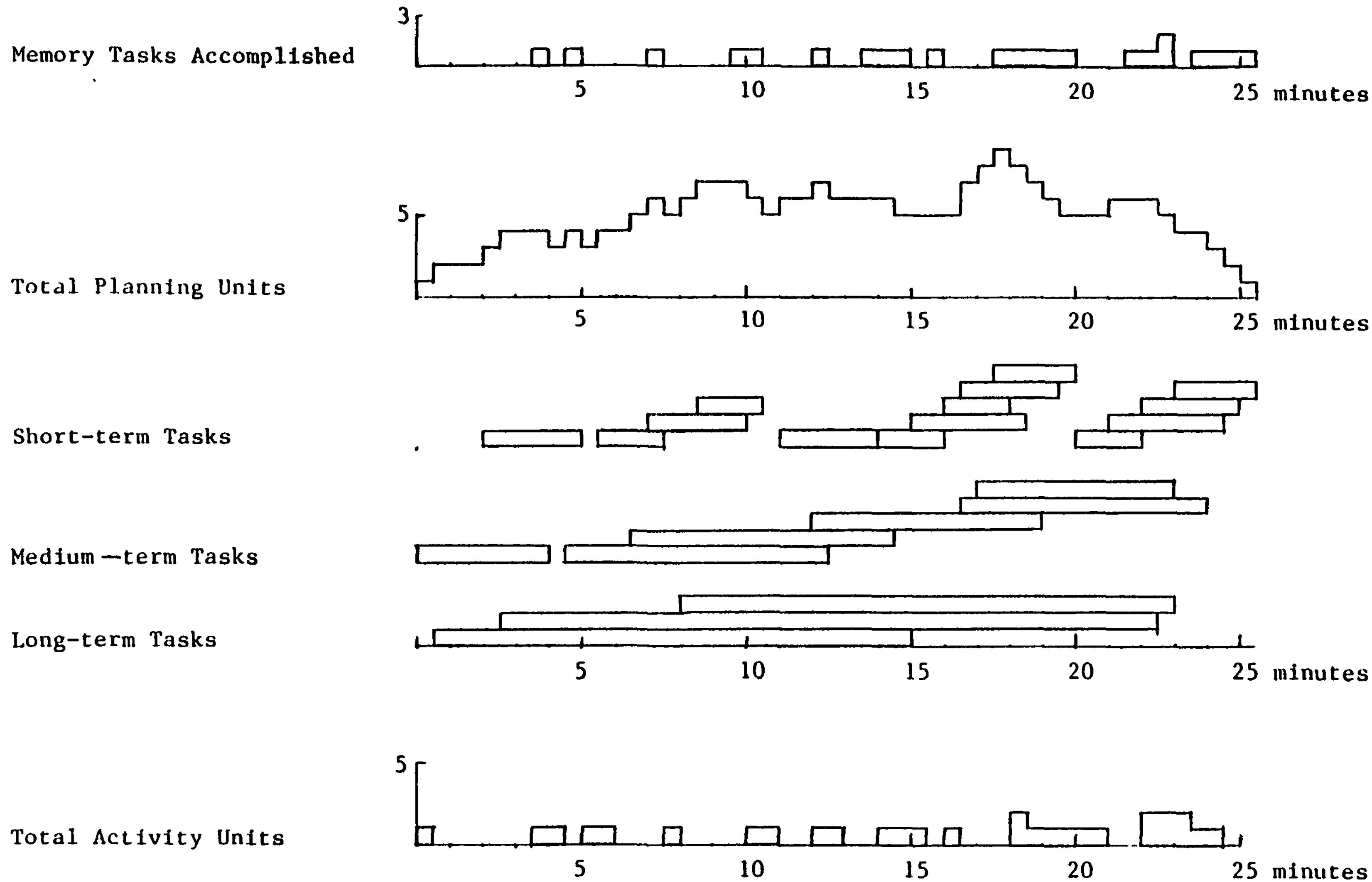


Figure 39: Planning scenario: Tasks and Workload Units versus elapsed time

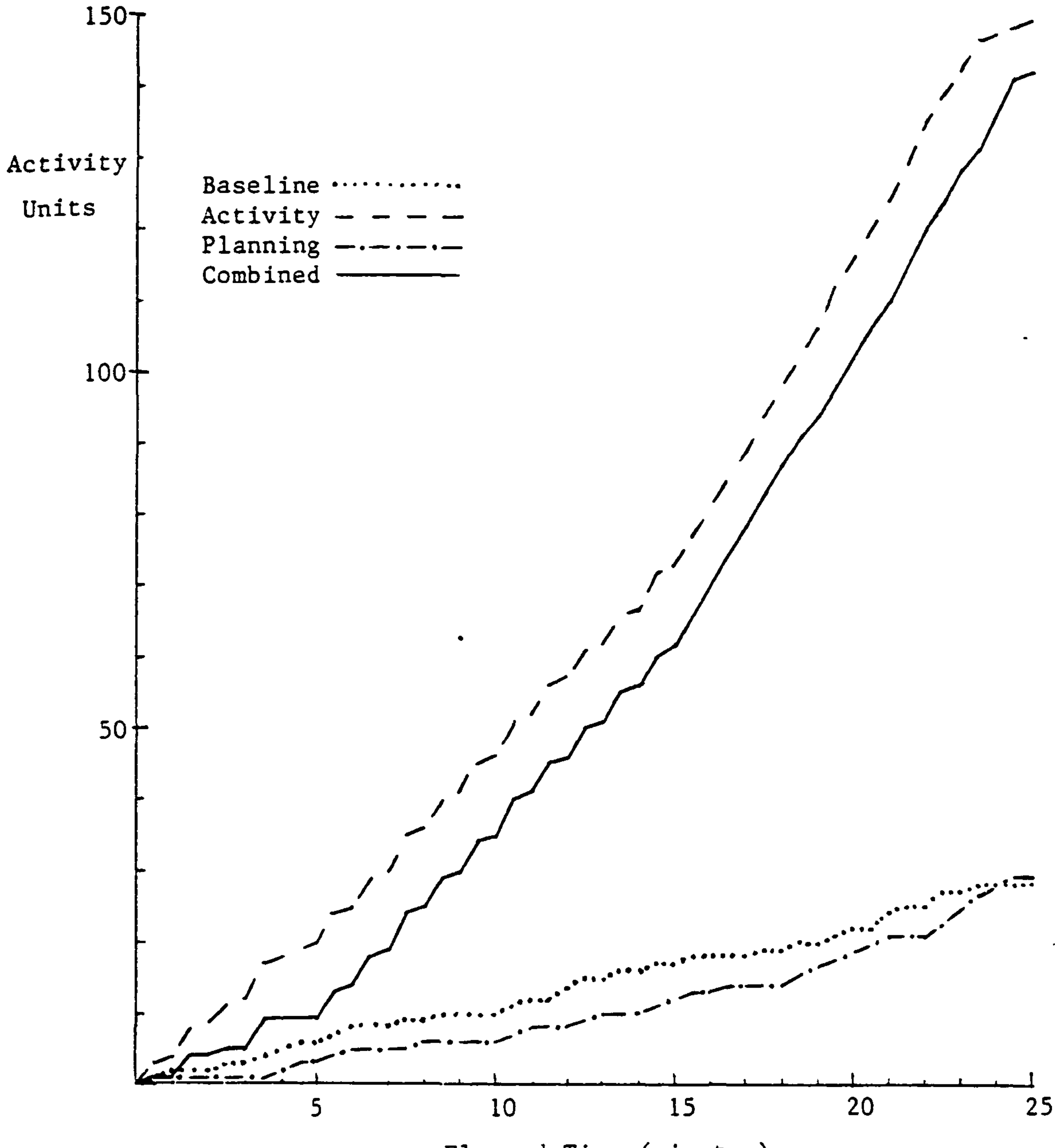


Figure 40: Accumulated Activity Workload Units versus elapsed time

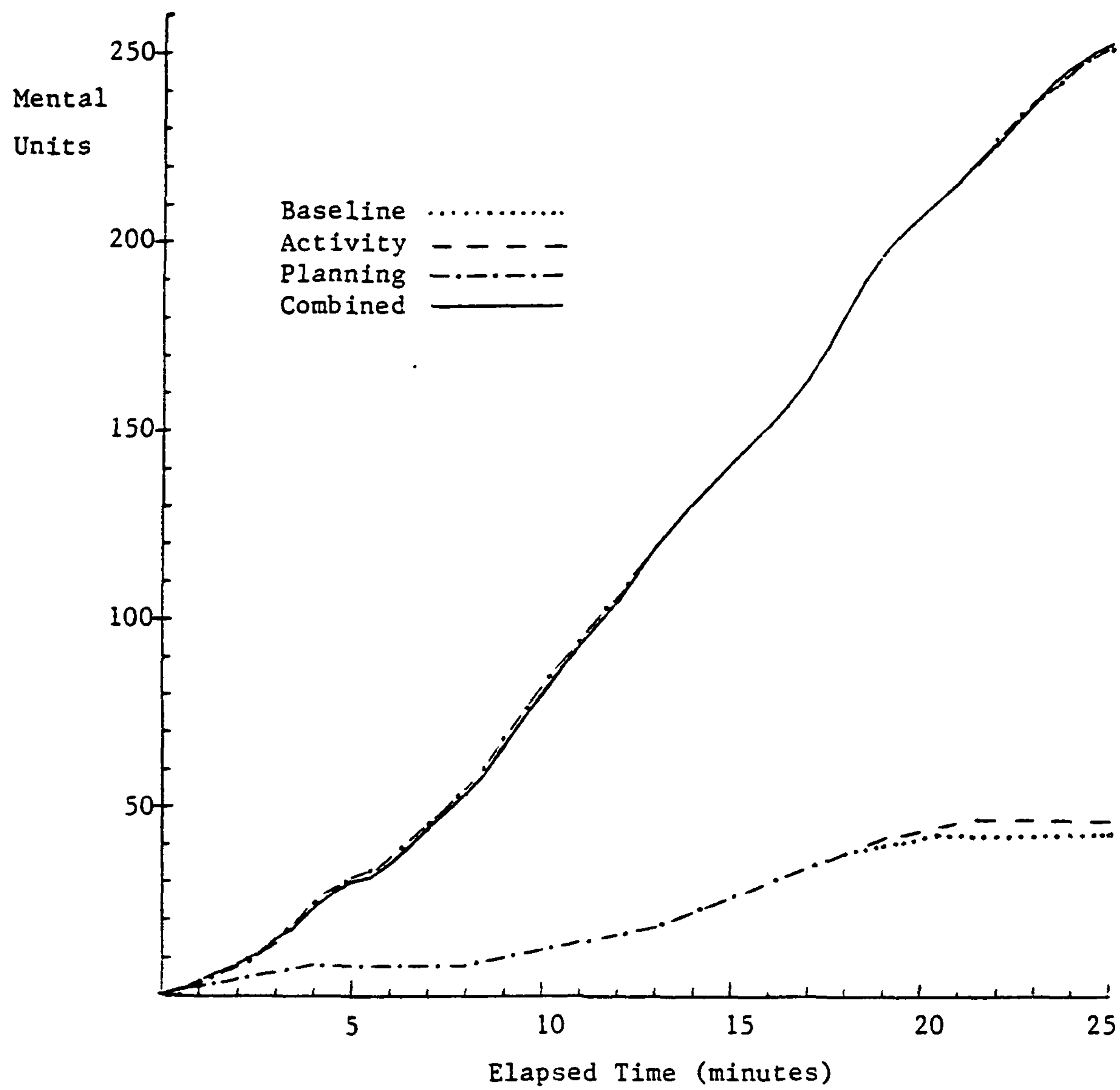


Figure 41: Accumulated Mental Workload Units versus elapsed time

Table 10: Scenario characteristics

	Scenario			
	Baseline	Activity	Planning	Combined
Total Planning WU's	43	47	253	254
Total Number of Planning Tasks	3	3	23	24
Short-term Planning Tasks	0	0	14	16
Medium-term Planning Tasks	3	3	6	5
Long-term Planning Tasks	0	0	3	3
Total Activity WU's	28	150	29	142

4.3 TRAINING AND INSTRUCTIONS

In addition to the initial screening sessions, each pilot participated in 4 to 10 hours of additional training. Three of the four pilots had flown the simulator before, but had never used the autopilot. They required about 4 hours of additional practice.

This autopilot is different from most commercial equipment. Longitudinal and Lateral modes must be engaged separately, adding one additional step to selecting some autopilot functions. (See Appendix 2 for a full description of the autopilot and autopilot dynamics.)

Those subjects who hadn't previously flown the simulator needed familiarization with several additional things. First, there was the flight instrument display. Two of the civilian pilots had only limited experience with an ADI/HSI presentation. The other two civilian pilots had no ADI/HSI background. Second, since all of the many switches and controls were mounted on the relatively small Control Box, it took a good deal of practice to become familiar with the switches, their functions, and their relative placement. Third, the dynamics and flight control sensitivity of a high-performance aircraft were new to the civilian pilots. Fourth, the control-stick or joy-stick was a completely new device for two of the civilian pilots. They had to become familiar with the problem of inadvertently coupling lateral and longitudinal inputs. Finally, the four pilots who hadn't flown the simulator before had to get used to the "feel" of the control-stick. The Control Box had rather

small springs for returning the control stick to a near neutral position. Thus, the pilot does not feel the magnitude of force feedback he is used to in most aircraft.

Before a session's data runs, pilots "warmed up" by flying instrument approaches, turns to headings, etc., for 20 to 30 minutes. After this warm up period, the pilots were handed the Instruction Sheet reproduced in Figure 42, the Subjective Ratings/Comments Sheet shown in Figure 43, and a Workload Reference Sheet (Figure 29).

In the instructions, pilots were told to fly "as well as you can" and follow all directions "to the best of your ability". They were also told that they would be scored on their ability to "follow instructions and comply with requests". Thus, they had no idea which parameter(s) would be measured. Any or all might be scored.

There was a distinct danger that the less experienced pilots would feel they were being compared directly with the more experienced pilots, and therefore feel additional stress. To combat this, all pilots were verbally assured that the primary focus of the experiment was the variations in their performance.

The Subjective Rating Sheet (Figure 43) was similar to, but different from, that used in the preliminary experiments. The greatest difference was in the number of divisions per scale. The preliminary experiment rating sheet had 10 divisions per scale. This new sheet had only two.

When using the rating sheet of Figure 28, the pilots often placed their ratings neatly between divisions or directly at the divisions. They did this despite instructions to consider the scale as being continuous. Therefore, to eliminate or reduce this "digitizing" effect, the only internal division in the new sheet's scales was at the half-way point. In theory, this provided one additional anchor while giving the subconscious a greater role in placing the ratings at an appropriate point.

In addition, scale descriptors on the new Rating Sheet were changed from "low" and "high" to something more appropriate to the individual scale. The Instruction Sheet also gave a short description of what was intended by each of the five measures.

As explained in the instructions, the simulation was "frozen" for subjective ratings at 5:00, 16:00, and 27:00 minutes elapsed time. The instructions explained the desired scoring style and noted that one minute was allowed for making the ratings during each break. The preliminary experiment had shown that the pilots only required about 20 to 30 seconds for making their ratings.

After each run, the pilots were debriefed and asked to put any

Instructions

This experiment investigates pilot workload. It is funded by NASA's Ames Research Center and the results will be incorporated in a forthcoming report.

Each session will consist of two flights. Each flight follows a similar ground track to an ILS approach to Runway 36, as is shown on the Navigational Chart.

Your task is to fly as well as you can, following all directions and requests to the best of your ability. You will be scored on your ability to follow instructions and comply with requests. Some requests/directions will be related to a geographic point, such as VOR #1. Most will be linked to a specific indicated elapsed time. This indicated elapsed time is displayed on the instrument panel directly above the CWS display and below the airspeed display. Please note that when you are not being vectored, you are expected to lead your turns when leaving one course and intercepting another.

There will be a number of memory or planning tasks such as "Climb to 3000 at 20:00". The 20:00 refers to the indicated elapsed time. There may be many such tasks, or only a few. There may be an overlap in the tasks. To help you plan, remember, and execute these tasks, it is suggested that you write down these requests.

At 05:00, 16:00, and 27:00 indicated elapsed time, the simulator will be "frozen" and you will make five subjective interpretations about how hard you are working or difficult the task is. The categories are:

Activity-Level: How busy are you? Are you bored or nearly as active as you can be?

Complexity: How complicated is the scenario, the required actions, or the planning required?

Difficulty: How tough is your task?

Stress: Do you feel pressured?

Workload: See the accompanying sheet for explanations.

As shown in the example below, please make your first rating with a 1, your second with a 2, and your third with a 3. You will have one minute during each break to make all 5 ratings.

2 3 1

After the flight is over, you will have an opportunity for comments and explanations. Please feel free to ask any questions you may have.

Thank you for your time and effort.

Figure 42: Primary Experiment Pilot Instructions

Activity-Level

bored continuous activity

Complexity

simple complicated

Difficulty

easy very hard

Stress

none high

Workload

light very heavy

Activity-Level

bored continuous activity

Complexity

simple complicated

Difficulty

easy very hard

Stress

none high

Workload

light very heavy

Figure 43: Subjective Ratings/Comments Sheet for Primary Experiment

comments or explanations on the rear of the Rating Sheet.

4.4 DATA

For this set of experiments, the computer was reprogrammed. It no longer recorded Control Box manipulations, but now recorded airspeed every 10 seconds in addition to the aircraft's x, y, and z position. As described in section 3.4, this data yielded a ground track, and by comparing position and elapsed time, desired altitudes and airspeeds were determined. This information was then compared with the actual airspeeds and altitudes to derive altitude and airspeed error. Altitude errors were not computed during directed climbs and descents and airspeed errors were not computed during directed airspeed changes. Pilots were expected to climb or descend at a minimum of 1000 feet per minute and accelerate or decelerate to the desired airspeed within 30 seconds or at a rate of at least 50 knots per minute for airspeed changes greater than 25 knots. These rates of change are consisted with recommended piloting techniques.

As explained in section 3.4, ground tracks were plotted for reference but deviations from the nominal ground track were not scored.

Altitude deviations still seemed to be the "best" objective measure. However, with only one objective measure, it was possible that pilots might give higher priority to one aspect of aircraft control than another. If altitude control improved, was airspeed control deteriorating? If altitude control deteriorated, was airspeed control improving? Measuring only one variable would miss this trade-off. Thus, airspeed deviations were scored to serve as a check of this possibility. Both variables were scored using mean absolute and RMS deviations.

Just as in the preliminary experiments, Subjective ratings were made for Activity Level, Complexity, Difficulty, Stress, and Workload. Ratings were made at three points during each run. However, this time the subjects were not asked to make an overall rating after each run. This was because the "overall" ratings made during the preliminary experiments were nearly identical to the arithmetic mean of the three segment ratings.

The distance from the left edge of each scale to each pilot rating was measured, divided by the total scale length, and multiplied by ten. This resulted in subjective ratings with a possible range of 0 to 10, just as for the preliminary experiment.

Again, five experimentally proven subjective ratings were used in order to examine the multi-dimensionality of the mental workload.

An integral aspect of this set of experiments was an investigation into not only the degree of mental workload, but also the effect this effort had on observable pilot behavior. Thus, in addition to the aircraft control measures and subjective ratings just discussed, other aspects of pilot behavior were also measured.

As explained in section 4.2, each scenario had a certain number of planning or memory tasks. During each run, notes were made on each pilot's compliance in carrying out these assigned tasks. This gave information on short-term, medium-term, and long-term tasks as well as "positional" versus "non-positional" memory tasks. (See Section 3.5) All pilots were assigned specific elapsed times (clearly displayed on the instrument panel) at which to perform these tasks. Each pilot was given $\pm$ 15 seconds from the designated time in which to begin the task. If a task was accomplished outside these limits, it was noted.

When a task was performed improperly, for example climbing to a wrong altitude or accelerating 10 knots instead of climbing 1000 feet, this was also noted.

A third type of mental error was forgetting or missing an item entirely.

A final source of information was post run debriefings. The pilots had many interesting and useful insights into mental workload, stress, and the ways these affected performance.

Chapter 5

PRIMARY EXPERIMENT: RESULTS

5.1 TRAINING AND LEARNING EFFECTS

Section 4.2 explained in detail how the experiment was designed to minimize "learning effect". First, the pilots flew the scenarios in different orders so that the runs were roughly counterbalanced. However, the counterbalancing was compromised to minimize the probability of some pilots being hopelessly overwhelmed. Thus, the Combined scenario was never flown first. This was the only concession to counterbalancing.

Second, as mentioned briefly in Chapter 4, each session's Baseline run acted as a "warm up" run and served as a day-to-day metric for the Subjective ratings. For each Subjective rating, the Baseline run ratings were averaged across all seven pilots and all three runs for each pilot. This yielded an overall mean baseline rating. This mean rating was added to the difference of a session's Baseline rating and second run (Activity, Planning, or Combined) rating. This gave an "adjusted" second run rating. The intent was to compensate for day-to-day differences in emotional state, stress, fatigue, et cetera.

Third, each subject was a highly trained pilot, went through a rigorous screening process, and was then trained on the simulator for an additional 5 to 15 hours. At the end of this training period, the pilots appeared to have passed the "knee" of the performance curve.

Katz [10] conducted a similar experiment and had considerable difficulty. He warned that a major problem was encountering a marked learning curve "...despite concerted efforts to circumvent learning curve effects by establishing a rather long briefing/warmup flight period."

He found that, "Performance stabilization and verbal questionnaires are inadequate indicators of learning curve plateaus." In the case of mental workload, these traditional indicators are not sufficient. Because of the "accumulator effect", subjects may show excellent performance but assess lower and lower workload ratings. Referring to Figure 4, as subjects' experience and expertise increase, they can maintain constant performance even though their subjective workload decreases.

So, the Objective and Subjective data was examined for

"learning effects". Using Student t-test and F-test techniques, I found a learning effect at a 90 percent confidence level for altitude deviations in the Baseline scenario. However, there was no significant learning effect in the altitude deviations for the Activity, Planning, or Combined scenarios. The Baseline altitude deviations were small, and did not enter into further results analysis.

There was no sign of learning effect in the airspeed deviation data for any of the four scenarios.

An examination of the Subjective Rating data yielded mixed results. Using the adjusted ratings, there was no "learning effect" for any of the ratings for the Activity scenarios. For the Planning scenario, only the Workload ratings showed a weak (80 percent confidence level) learning effect.

The extensive training, the modified counterbalancing of scenarios and subjects, and "adjusting" the subjective ratings appears to have minimized learning effect for the Activity and Planning scenarios.

However, there was some evidence of learning effect for the Combined scenario. Three subjective ratings were lower for the third sessions than the second sessions. The effect was at an 80. percent confidence level for Complexity ratings. Since post-run debriefings showed that Complexity ratings were closely tied to the pilots' ease with the autopilot, this may be due to greater familiarity with the device. Learning effect was at a much stronger 95 percent confidence level for the Difficulty and Workload ratings. This is understandable. None of the practice rounds were nearly as intense as the Combined scenario. Furthermore, the Combined scenario was the sum of the Activity and Planning scenarios. Subjects who had seen both the Activity and Planning scenarios before flying the Combined scenario had an advantage over those who flew the Combined scenario after flying only one of the others.

Finally, an analysis of variance showed no statistically significant difference for planning task performance for any scenario.

5.2 ALTITUDE AND AIRSPEED DEVIATION RESULTS

Altitude error data was synthesized from the computer's output and is shown in Table 11. Both mean absolute error and rms error data is listed for each pilot, scenario, and segment. This data, in turn, is consolidated into overall error data for scenarios and segments in Table 12.

Table 11: Individual mean absolute and rms altitude deviations (feet)

SCENARIO		Pilot						
		A	B	C	D	E	F	G
Baseline								
Segment I:	mean	34.9	33.9	45.4	76.9	29.4	31.1	18.2
	rms	41.6	37.8	54.7	85.3	30.4	44.3	20.1
Segment II:		42.5	38.5	87.5	48.8	29.6	26.6	11.0
		49.5	43.4	93.4	53.9	36.0	35.3	14.6
Segment III:		26.1	26.1	56.0	25.4	33.2	22.8	14.5
		29.3	28.8	67.9	27.0	56.2	30.6	16.3
Overall:		34.5	32.8	63.6	50.4	30.7	26.8	14.6
		40.1	36.7	72.0	55.4	40.9	36.7	17.0
Activity								
Segment I:	mean	71.6	46.3	342.1	172.0	84.8	52.9	31.2
	rms	78.7	49.0	420.2	273.4	110.5	67.6	35.4
Segment II:		111.8	93.8	128.3	101.1	86.8	111.5	50.7
		165.7	141.4	205.2	140.1	111.0	242.6	67.6
Segment III:		106.0	143.6	163.2	147.8	111.9	198.8	94.5
		172.0	201.0	233.1	253.3	128.0	272.6	133.2
Overall:		96.5	94.6	211.2	140.3	94.5	121.1	58.8
		138.8	130.5	286.2	222.3	116.5	194.2	78.7
Planning								
Segment I:	mean	5.8	71.2	11.9	14.0	7.0	13.7	12.9
	rms	9.3	93.4	12.7	15.1	7.2	14.0	13.1
Segment II:		57.0	55.5	59.8	67.7	41.9	43.2	9.7
		66.7	72.8	60.3	80.6	53.7	48.0	13.8
Segment III:		44.4	61.6	110.6	91.6	27.2	27.2	26.4
		62.0	66.8	110.6	95.4	28.9	29.0	30.1
Overall:		35.7	62.8	60.8	57.8	25.4	28.0	16.3
		46.0	77.7	61.2	63.7	29.9	30.3	19.0

Table 11, continued

SCENARIO		Pilot						
		A	B	C	D	E	F	G
Combined								
Segment I:	mean	109.3	23.1	176.7	230.4	89.4	8.3	17.1
	rms	139.6	33.0	231.3	375.4	115.5	9.2	18.0
Segment II:		191.9	58.0	249.6	157.1	53.9	83.4	63.2
		292.7	117.8	424.0	220.9	97.8	145.8	88.7
Segment III:		313.0	72.2	195.0	167.9	137.5	101.6	96.3
		405.4	107.8	269.5	210.4	176.3	136.5	128.6
Overall:		204.7	51.1	207.1	185.1	93.6	64.4	58.9
		279.2	86.2	308.3	268.9	129.9	97.2	78.4

Table 12: Overall mean absolute and rms altitude deviations (feet)

SCENARIO	SEGMENT	MEAN	STD DEV	RMS
Baseline	I	39.1	18.7	50.6
	II	41.4	24.0	51.0
	III	30.6	13.8	41.4
	Overall	37.0	19.0	47.7
Activity	I	114.4	110.5	147.8
	II	97.7	24.8	153.3
	III	138.0	36.6	199.0
	Overall	116.7	67.3	166.7
Planning	I	19.5	23.0	23.5
	II	47.8	19.1	56.6
	III	55.6	34.0	60.4
	Overall	41.0	29.5	46.8
Combined	I	93.5	85.6	131.7
	II	122.4	77.5	198.2
	III	154.8	81.8	204.9
	Overall	123.6	81.7	178.3

Note the standard deviation data in Table 12. Comparing these standard deviations with the individual pilot performance data from Table 11, one can see that the bulk of pilot deviations tended to lie near the mean. However, there was usually some pilot whose deviations took an extreme, isolated jump, inflating the standard deviation for the group.

In general, just as the WU rate increased from Segment I to Segment III, so did altitude deviations. (see Figure 44) The exaggerated effect of large deviations in the rms data concealed any statistically significant segment-to-segment differences for the Baseline or Activity scenarios, but the mean absolute error data yielded significant differences for all four scenarios. Using F-tests, segment-to-segment mean error differences were significant at a 90 percent confidence level for the Combined scenario, 95 percent for the Baseline and Activity scenarios, and 99 percent for the Planning scenario. The larger spread of individual performance in the Combined scenario was responsible for its lower confidence level result.

Differences and similarities among the scenarios were striking. First of all, the magnitude of altitude deviations was a strong function of the mode of aircraft control. As Figure 44 shows, there was a considerable difference between the manually controlled Combined and Activity scenarios and the autopilot controlled Planning and Baseline scenarios. Using a t-test, the difference was significant with 99 percent confidence. The average deviation was 3.1 times greater (120.2 feet versus 39.0 feet) under manual control, and the rms deviation was 3.6 times greater (172.5 feet versus 47.3 feet). However, all the difference cannot be ascribed simply to manual control versus autopilot control differences. The manually controlled Combined and Activity scenarios also had much more difficult altitude profiles than the autopilot controlled scenarios. (See Figure 37)

Interestingly, the magnitude of mental tasking had no significant impact on the magnitude of the altitude deviations. The Baseline scenario's altitude deviations were statistically similar to those of the Planning scenario which differed from it solely in having a large number of mental planning tasks. Similarly, the Activity and mentally demanding Combined scenario results were statistically identical.

Airspeed error data was also synthesized from the computer's output, and is presented in Tables 13 and 14. Like the altitude deviation data, some of the large standard deviations in Table 14 are due to some pilot's momentary lapse. Most of the deviation data was fairly consistent in magnitude.

Unlike the altitude deviation data, segment-to-segment differences in both mean absolute and rms airspeed errors were

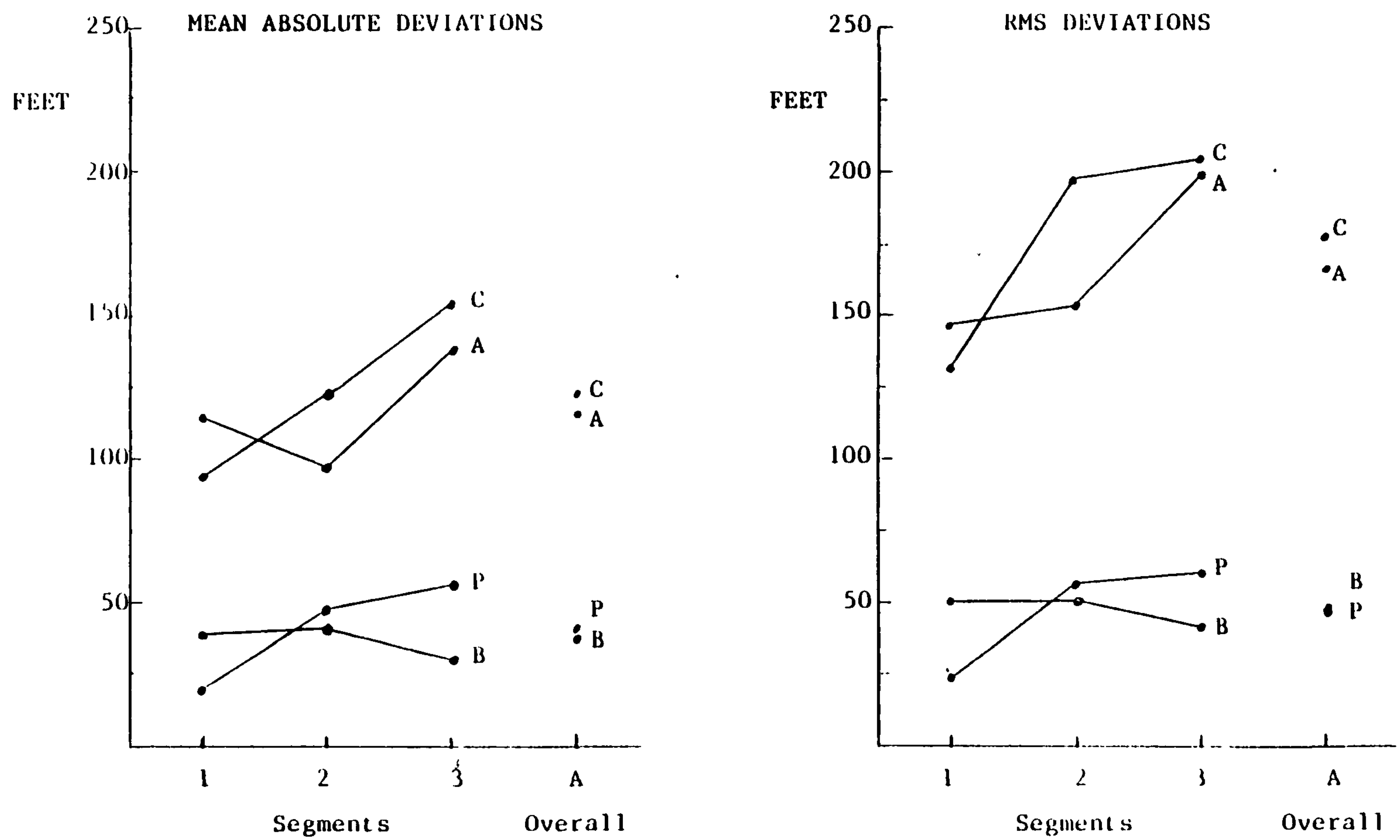


Figure 44: Average altitude deviations for the Baseline (B), Activity (A), Planning (P), and Combined (C) scenarios

Table 13: Individual mean absolute and rms airspeed deviations (knots)

SCENARIO		Pilot						
		A	B	C	D	E	F	G
Baseline								
Segment I:	mean	3.1	1.7	1.7	1.7	2.8	1.1	1.3
	rms	3.5	2.6	2.0	2.6	3.2	1.4	2.0
Segment II:		4.0	3.5	3.5	4.0	5.0	2.9	4.5
		4.1	5.5	4.2	4.4	5.6	3.7	4.8
Segment III:		3.0	2.5	3.6	3.9	7.1	2.2	1.3
		3.3	2.9	4.3	4.7	8.9	2.4	1.7
Overall:		3.4	2.6	2.9	3.2	5.0	2.1	2.4
		3.6	3.7	3.5	3.9	5.9	2.5	2.8
Activity								
Segment I:	mean	7.8	4.9	7.7	5.9	22.1	3.6	2.9
	rms	8.5	5.5	8.8	7.4	25.3	4.7	3.9
Segment II:		17.4	10.0	11.7	8.3	9.6	3.9	5.9
		22.4	13.4	14.3	12.9	12.6	4.9	7.1
Segment III:		22.4	18.0	8.8	8.9	10.7	7.2	7.4
		30.9	22.2	10.3	11.4	13.3	10.4	10.1
Overall:		15.9	11.0	9.4	7.7	14.1	4.9	5.4
		20.6	13.7	11.1	10.6	17.1	6.7	7.1
Planning								
Segment I:	mean	0.2	1.3	1.2	0.2	0.8	0.7	0.6
	rms	0.3	2.4	2.1	0.3	0.8	0.7	0.6
Segment II:		7.5	3.7	4.1	2.4	0.8	1.5	6.1
		7.5	4.2	4.2	3.0	0.9	1.7	6.3
Segment III:		1.9	3.2	7.3	1.9	1.7	3.4	3.6
		2.5	4.2	7.3	2.1	2.0	5.3	3.8
Overall:		3.2	2.7	4.2	1.5	1.1	1.9	3.4
		3.4	3.6	4.5	1.8	1.2	2.6	3.5

Table 13, continued

SCENARIO		Pilot						
		A	B	C	D	E	F	G
Combined								
Segment I:	mean	6.7	4.6	9.3	6.0	3.3	2.3	4.1
	rms	8.1	4.9	10.6	8.2	4.1	3.1	4.3
Segment II:		15.3	9.9	17.4	10.4	12.6	6.2	5.1
		18.9	13.3	22.8	14.7	16.8	7.3	6.1
Segment III:		8.0	9.0	9.7	11.4	17.9	5.8	5.4
		10.7	11.1	12.4	17.2	25.3	8.7	7.1
Overall:		10.0	7.8	12.1	9.3	11.3	4.8	4.9
		12.6	9.8	15.3	13.4	15.4	6.4	5.8

Table 14: Overall mean absolute and rms airspeed deviations (knots)

SCENARIO	SEGMENT	MEAN	STD DEV	RMS
Baseline	I	1.9	0.7	2.9
	II	3.9	0.7	5.0
	III	3.4	1.9	4.4
	Overall	3.1	1.4	4.1
Activity	I	7.9	6.6	9.2
	II	9.5	4.3	12.5
	III	11.9	5.9	15.5
	Overall	9.8	5.7	12.4
Planning	I	0.7	0.4	1.0
	II	3.7	2.4	4.0
	III	3.3	1.9	3.9
	Overall	2.6	2.2	3.0
Combined	I	5.2	2.4	6.2
	II	11.0	4.5	14.3
	III	9.6	4.2	13.2
	Overall	8.6	4.4	11.2

significant for all four scenarios (See Figure 45). For mean absolute airspeed errors, the segments differed at a 90 percent confidence level for the Activity scenario and a 99 percent level for the Baseline, Planning, and Combined scenarios. RMS airspeed errors differed at a 95 percent confidence level for the Baseline and Activity scenarios and a 99 percent confidence level for the Planning and Combined scenarios.

Like the altitude deviation data, the magnitude of airspeed errors was a strong function of the mode of aircraft control. As shown in Figure 45, when airspeed was under manual control, deviations were much greater than when airspeed was under autopilot control. The difference was statistically significant at a 99 percent confidence level for mean absolute error and a 98 percent level for rms errors. Mean absolute airspeed deviations were 3.3 times as large (9.2 knots to 2.8 knots) and rms deviations were 3.4 times as large (11.8 knots to 3.5 knots) when the simulator was flown manually rather than with the autopilot. Part of this result may be due to the much more difficult airspeed profile for the manually controlled scenarios (See Figure 35).

This airspeed deviation data also showed little mental tasking effect. There was no significant difference between scenarios which had similar manual activity levels but different planning workloads.

Both altitude and airspeed deviations were similar for all the pilots. In general, the low experience pilots had slightly higher deviations than the most experienced pilots. However, there was enough scatter in the data to keep the differences statistically insignificant.

This objective data showed only a hint of performance degradation due to pilot workload saturation. During the Activity scenario runs, only two pilots out of seven had average mean altitude deviations greater than 150 feet in Segment III, and two other pilots had average mean airspeed deviations greater than 15 knots in Segment III. For the Combined scenario, the number of saturated pilots rose to three for the altitude deviations and remained at 2 for the airspeed deviations.

Within each scenario, there was no significant correlation between airspeed and altitude deviations. This was due to some evidence of altitude and airspeed control trade-offs for various individuals during all four scenarios. However, overall airspeed and altitude control correlated at a 95 percent confidence level when considering the four scenarios together. The Baseline and Planning scenarios had low deviations for each score and the Activity and Combined scenarios had high deviations for both scores.

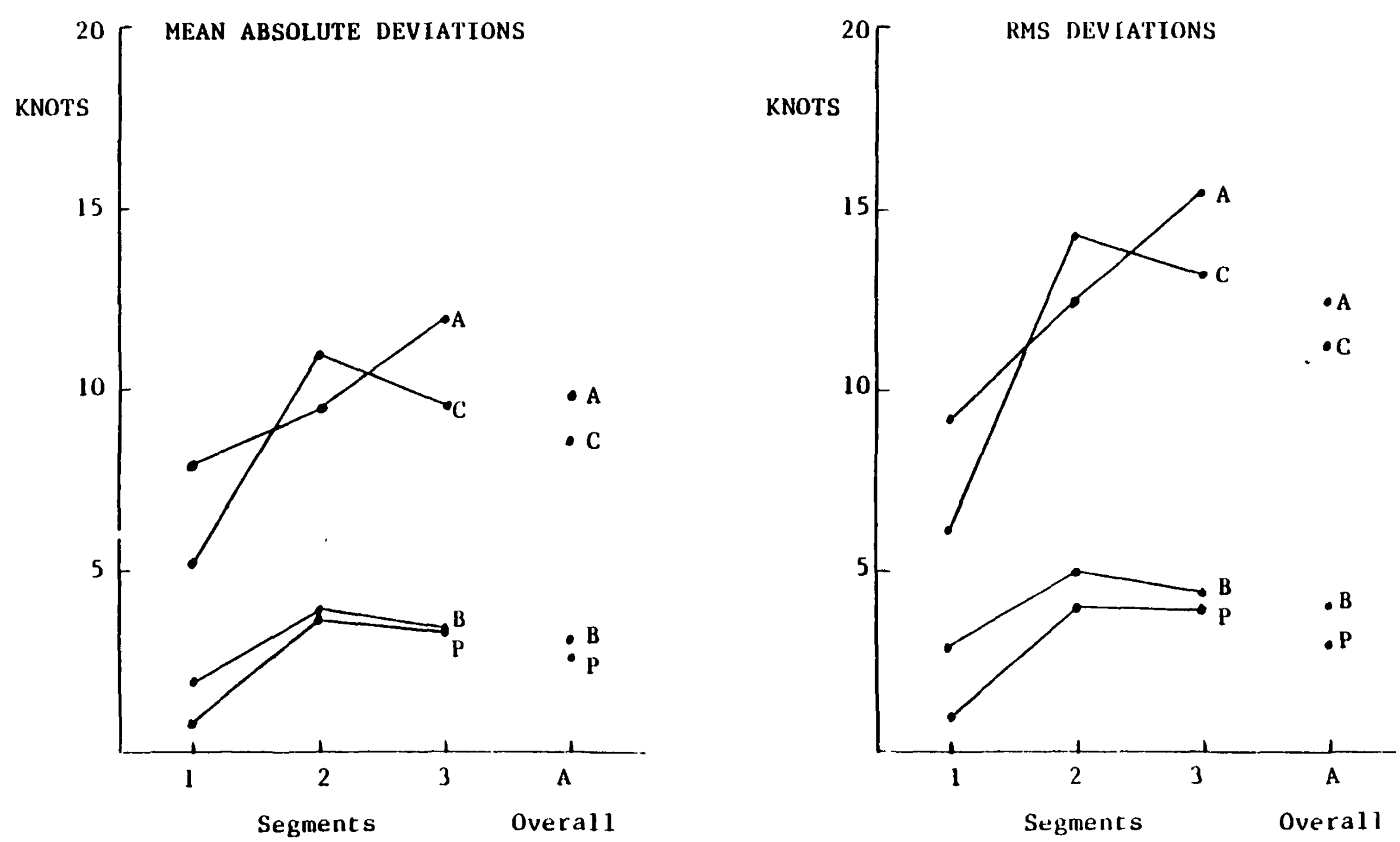


Figure 45: Average airspeed deviations for the Baseline (B), Activity (A), Planning (P), and Combined (C) scenarios

5.3 SUBJECTIVE RATING RESULTS

The Subjective Rating data was useful because it illustrated the impression these scenarios were making in the minds of the pilots. Thus, although only an indirect measure, one would expect these ratings to provide a better indication of mental workload than objective performance data.

Table 15 presents the average overall subjective rating for each pilot, scenario, and subjective category. Table 16 gives the ratings averaged over all the pilots for each segment, scenario, and category. Note that the standard deviation data in Table 16 is very consistent from rating to rating and scenario to scenario. It did not exhibit the wide variations present in the altitude and airspeed deviation data.

First, how well did each of the five ratings distinguish between the predominantly manual control oriented Activity scenario, the predominantly mental workload oriented Planning scenario, and the Combined scenario? ACTIVITY LEVEL ratings did not reliably distinguish between the Activity and Planning scenarios. The pilots apparently felt equally active in both, although the activities were of a fundamentally different nature. However, the difference between the Combined scenario and both of the others was significant at a 98 percent confidence level. Overall ratings are plotted in Figure 46.

COMPLEXITY rating results were similar. There was no significant difference between the Activity and Planning scenarios, but there was a difference between these scenarios and the Combined scenario at a 99 percent confidence level. See Figure 47.

DIFFICULTY ratings were somewhat different. As shown in Figure 48, the Activity scenario was rated slightly more difficult (80 percent confidence level) than the Planning scenario. The Combined scenario was considered more difficult than the Activity scenario (90 percent confidence) and definitely more difficult than the Planning scenario (99.9 percent confidence). Because the Activity and Combined scenarios were similar in all other respects, this difference in difficulty level is solely due to an added <u>planning</u> workload.

STRESS ratings indicated the Activity scenario was slightly more stressful than the Planning scenario (80 percent confidence level). However, as can be seen in Figure 49, the Combined scenario was definitely more stressful than either of the other two scenarios (99 percent confidence).

Finally, WORKLOAD ratings for the Activity and Planning scenarios were similar. The Combined scenario had a higher workload than the Activity scenario (90 percent confidence) or the Planning

Table 15: Average Subjective Ratings for each pilot (Adjusted)

	Pilot						
	A	B	C	D	E	F	G
BASELINE							
Activity Level	5.0	2.7	2.1	2.7	2.8	2.8	2.9
Complexity	4.7	2.4	1.6	3.5	2.6	2.9	2.1
Difficulty	4.0	3.1	1.4	2.6	2.5	2.9	2.1
Stress	3.7	2.3	1.8	2.4	2.5	2.6	1.3
Workload	2.4	2.4	1.2	2.3	2.4	2.8	2.1
ACTIVITY							
Activity Level	7.6	7.3	4.4	6.6	7.8	5.8	6.1
Complexity	3.7	4.1	3.8	4.9	7.2	5.6	3.5
Difficulty	6.4	6.9	3.7	6.0	6.5	6.0	4.7
Stress	4.1	5.4	3.8	5.3	6.5	5.5	3.5
Workload	6.1	5.4	2.5	5.7	6.8	6.4	5.3
PLANNING							
Activity Level	3.2	5.6	7.6	5.6	4.4	5.3	6.1
Complexity	3.2	4.5	7.2	4.6	4.6	4.6	5.4
Difficulty	2.9	4.4	6.5	3.9	4.5	4.4	5.3
Stress	3.2	5.6	2.4	5.4	4.6	3.2	4.7
Workload	4.1	6.4	7.0	4.2	4.4	4.2	4.0
COMBINED							
Activity Level	9.1	9.0	8.3	7.1	8.8	7.4	6.3
Complexity	6.9	5.7	10.1	5.7	7.6	7.1	5.5
Difficulty	5.7	8.4	10.7	7.3	8.1	6.6	6.4
Stress	8.1	8.2	7.5	7.6	8.9	5.9	5.2
Workload	7.1	8.4	10.3	7.6	8.8	6.2	5.5

Table 16: Average Subjective Ratings for each Segment (Adjusted)

SCENARIO	SEGMENT I	SEGMENT II	SEGMENT III	Overall	Std Dev
BASELINE					
Activity Level	2.6	2.8	3.5	3.0	0.9
Complexity	2.3	2.5	3.4	2.7	1.0
Difficulty	2.2	2.4	3.1	2.6	0.8
Stress	2.0	2.1	3.0	2.4	0.7
Workload	1.8	2.2	2.8	2.3	0.5
ACTIVITY					
Activity Level	5.4	6.7	7.3	6.5	1.2
Complexity	3.4	5.0	5.7	4.7	1.3
Difficulty	4.5	6.0	6.7	5.7	1.1
Stress	3.7	4.9	6.1	4.9	1.1
Workload	3.9	5.5	7.0	5.5	1.4
PLANNING					
Activity Level	4.1	5.1	7.0	5.4	1.4
Complexity	4.1	4.6	5.9	4.8	1.3
Difficulty	3.3	4.0	6.3	4.6	1.1
Stress	3.3	3.9	5.3	4.2	1.2
Workload	3.9	4.7	6.2	4.9	1.2
COMBINED					
Activity Level	5.9	8.3	9.8	8.0	1.1
Complexity	5.4	6.9	8.5	6.9	1.6
Difficulty	5.9	7.8	9.1	7.6	1.7
Stress	5.5	7.6	8.9	7.3	1.3
Workload	5.7	7.7	9.6	7.7	1.6

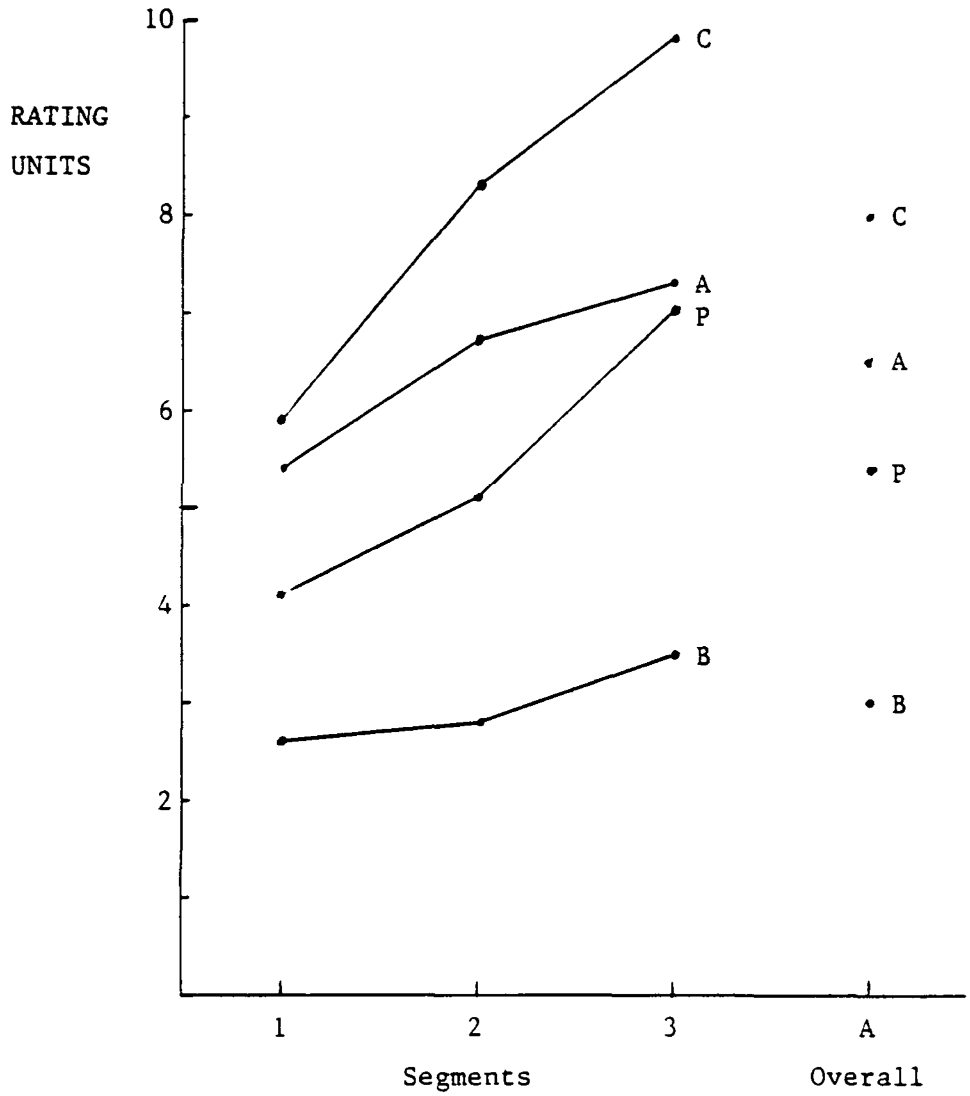


Figure 46: Average subjective ACTIVITY LEVEL ratings for the Baseline (B), Activity (A), Planning (P), and Combined (C) scenarios

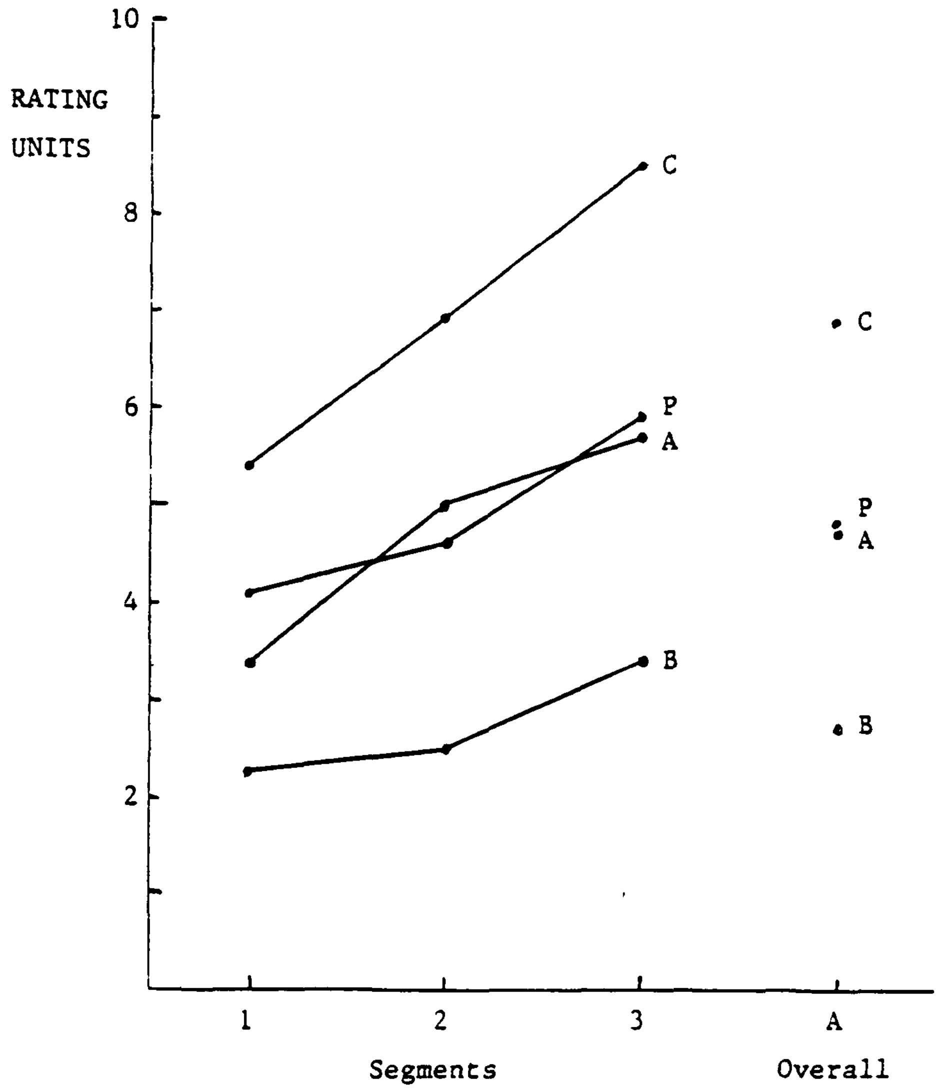


Figure 47: Average subjective COMPLEXITY ratings for the Baseline (B), Activity (A), Planning (P), and Combined (C) scenarios

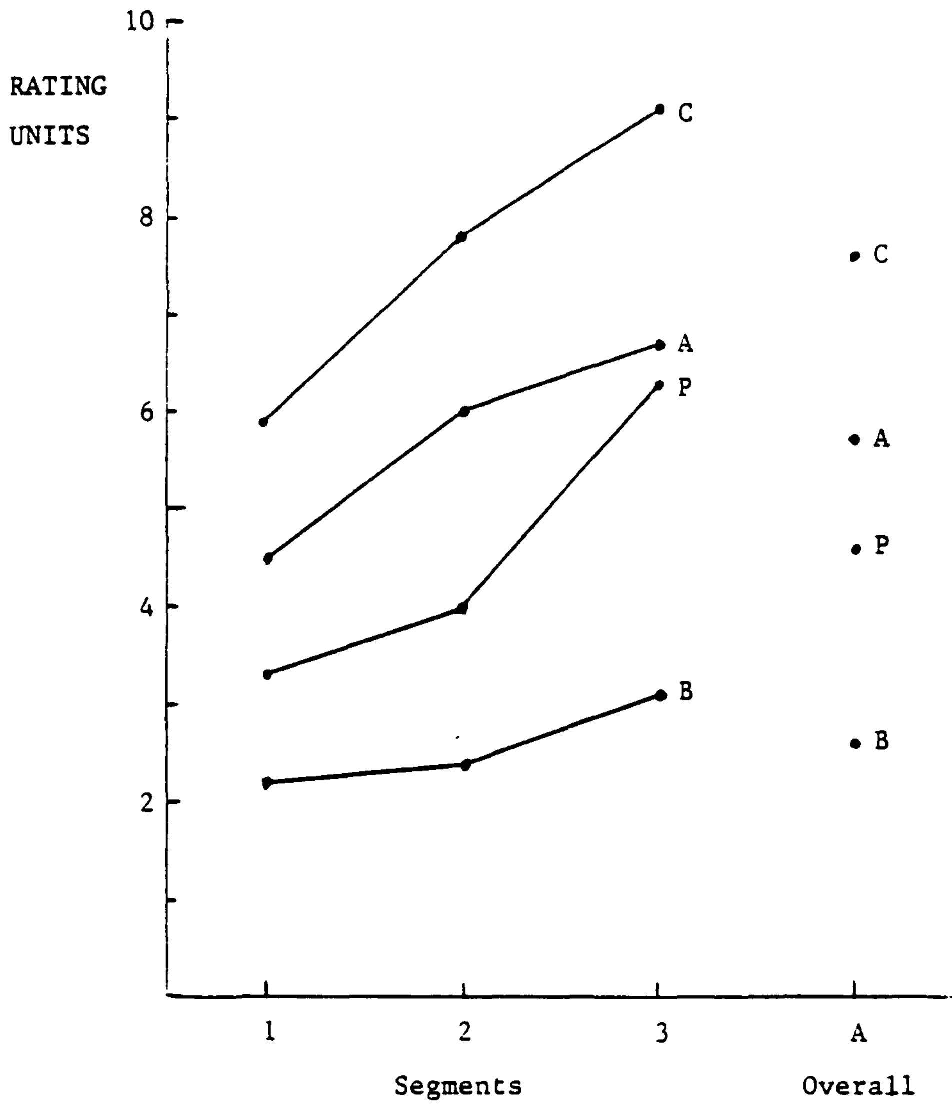


Figure 48: Average subjective DIFFICULTY ratings for the Baseline (B), Activity (A), Planning (P), and Combined (C) scenarios

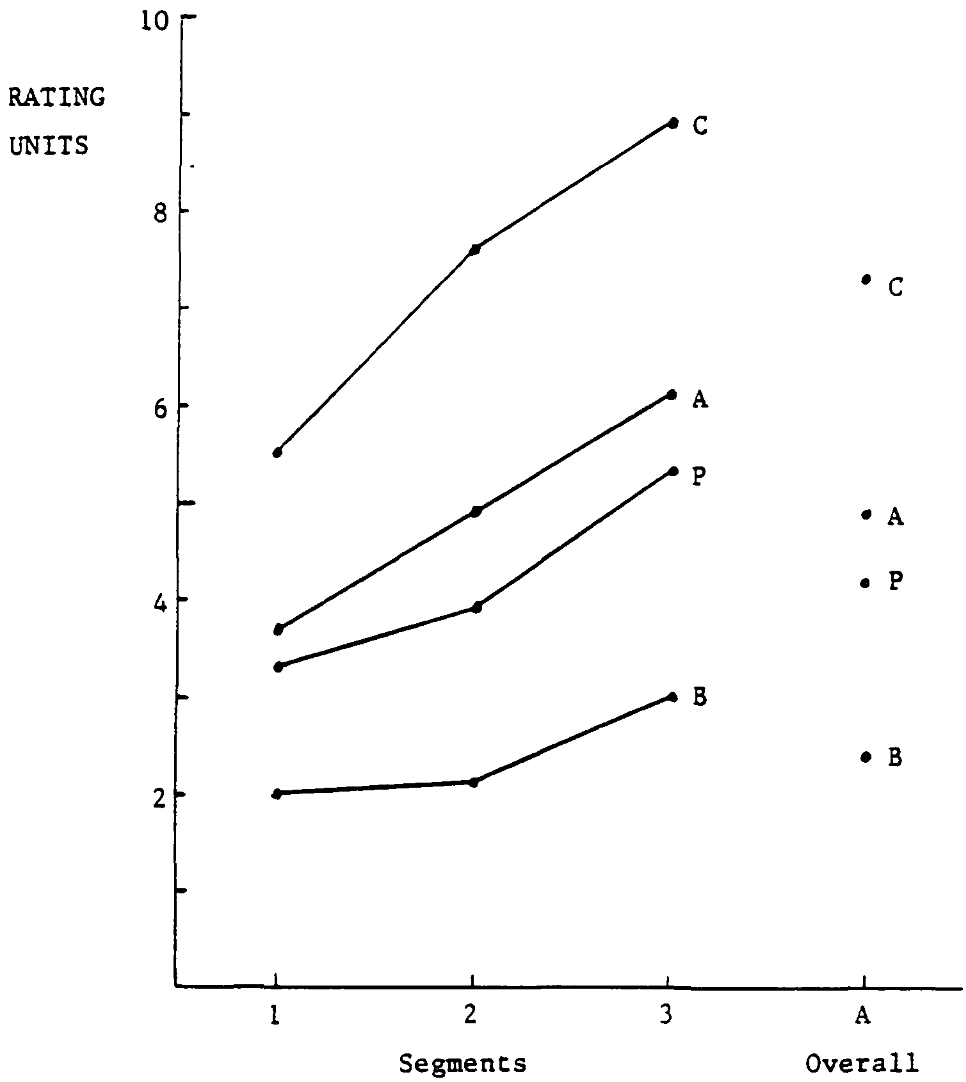


Figure 49: Average subjective STRESS ratings for the Baseline (B), Activity (A), Planning (P), and Combined (C) scenarios

scenario (99.9 percent confidence). See Figure 50.

The Planning scenario was essentially a Baseline scenario with an added mental workload component. The Activity scenario was a Baseline scenario complicated by a great deal of manual control work. The Combined scenario was a combination of the Activity and Planning scenarios. Therefore, the construction of the scenarios and the results plotted in Figures 46 to 50 led me to investigate whether this construct was reflected in the subjective ratings.

For all five ratings, I found the incremental difference between the Baseline scenario and each of the other three scenarios. I then examined how the sum of these increments for the Activity and Planning scenarios compared with the incremental Combined ratings. For example, suppose that the Baseline rating for Difficulty was 3.0 and the Difficulty ratings for the Activity, Planning, and Combined scenarios were 5.0, 5.3, and 7.5 respectively. The incremental ratings for the Activity, Planning, and Combined ratings would then be 2.0, 2.3, and 4.5. The sum of the Activity and Planning scenario increments would be 4.3. This increment (averaged with the increments for all the other pilot's increments) was compared with the Combined scenario's increment of 4.5 (averaged with the other pilot's Combined scenario increments).

The sum of the Activity and Planning increments for the Complexity, Difficulty, Stress, and Workload ratings was not statistically different from the incremental Combined ratings. The Activity Level rating did show a potential difference between the Combined rating and the sum, but at a low, 70 percent confidence level. The sum of the Activity Level ratings for the Activity and Planning scenarios is greater than the Combined scenario rating.

In view of the well established fact that the magnitude of subjective perception is logarithmically related to stimulus magnitude, this nearly linear response was somewhat surprising. At no point were the pilots ever told that the Combined scenario contained the sum of manual and mental tasks from the Activity and Planning scenarios.

Another item of interest was whether these five ratings differed from each other for each of the three non-Baseline scenarios. Table 17 lists confidence levels for a statistically significant difference between the ratings for the Activity scenario. The Activity Level ratings were different from the other four ratings. Complexity ratings differed significantly from Activity Level and Difficulty ratings. Difficulty ratings differed from Activity Level, Complexity, and Stress. Stress Ratings differed from Activity Level and Difficulty ratings. Workload ratings differed primarily from Activity Level ratings. Overall, the pilots found significant differences among these categories for this manual control type of activity.

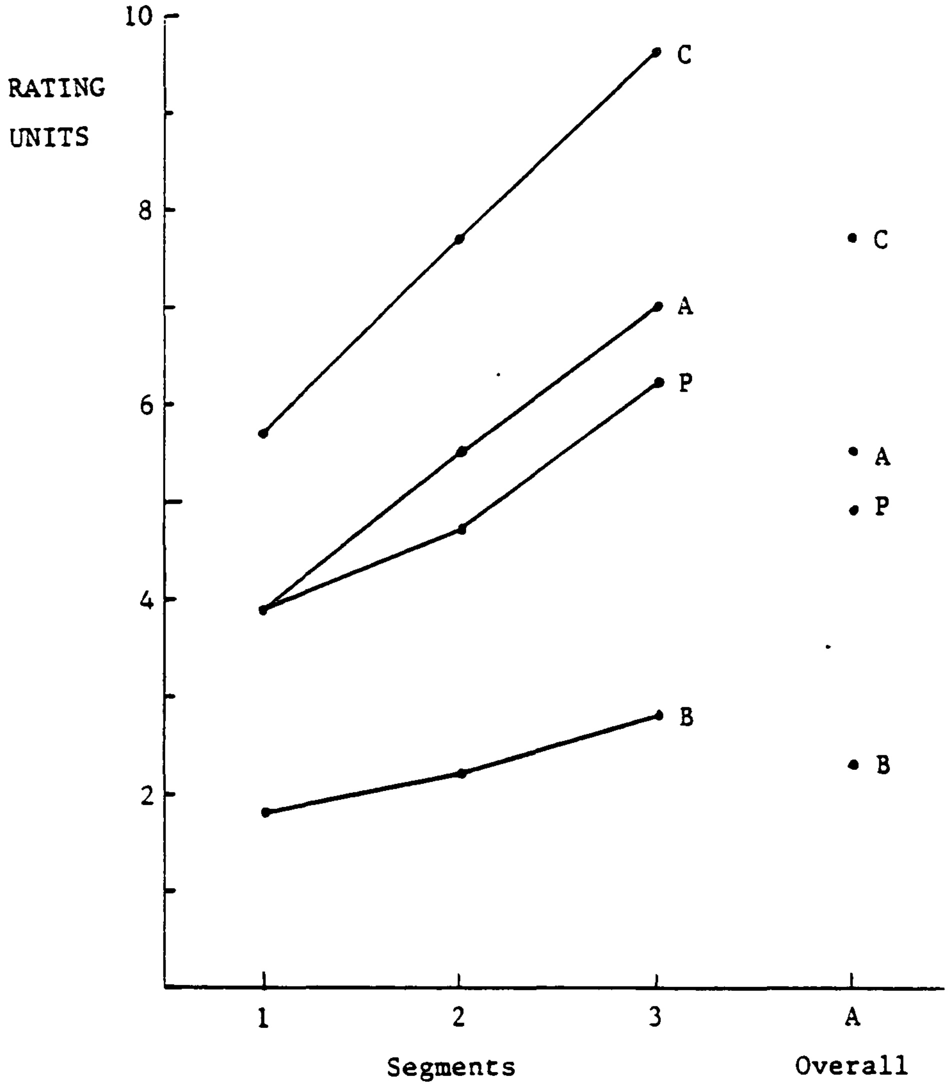


Figure 50: Average subjective WORKLOAD ratings for the Baseline (B), Activity (A), Planning (P), and Combined (C) scenarios

Table 17: Activity scenario: statistical confidence levels for differences between various subjective ratings

	Activity Level	Complexity	Difficulty	Stress	Workload
Activity Level	--	98	98	99	98
Complexity	98	--	90	40	80
Difficulty	98	90	--	95	60
Stress	99	40	95	--	70
Workload	98	80	60	70	--

The Stress and Complexity ratings were similar. Both were relatively low (Complexity: 4.7 average; Stress: 4.9 average). Some of the pilots commented that pure, intense manual activity was easier than pure, intense mental activity. This may explain why this scenario was rated the least "complex" of the three non-Baseline scenarios, and not very stressful.

Table 18 presents statistical difference data for the Planning scenario. Activity Level, Complexity, and Difficulty ratings all differed from each other. The Stress and Workload ratings were less distinct. For this mentally difficult scenario, the Stress ratings were similar to the Difficulty ratings. Workload ratings were similar to the Complexity ratings. This is consistent with pilot comments that most of their workload was related to the complex nature of the "sorting" and "planning" required for this scenario.

Table 18: Planning scenario: statistical confidence levels for differences between various subjective ratings

	Activity Level	Complexity	Difficulty	Stress	Workload
Activity Level	--	95	99	80	70
Complexity	95	--	98	60	20
Difficulty	99	98	--	20	60
Stress	80	60	20	--	60
Workload	70	20	60	60	--

Table 19 shows that there was little difference among the ratings for the Combined scenario. Only the Activity Level and Stress ratings were significantly different. Apparently, the subjects were successfully distinguishing between being busy and being stressed. Both Stress and Workload ratings were similar to the Difficulty rating. One possible explanation for these mostly negative results is that the manual and mental workload was so high that all of the ratings were high. Thus, perceptual differences couldn't manifest themselves in the statistics.

Table 19: Combined scenario: statistical confidence levels for differences between various subjective ratings

	Activity Level	Complexity	Difficulty	Stress	Workload
Activity Level	--	80	40	95	40
Complexity	80	--	70	40	80
Difficulty	40	70	--	20	20
Stress	95	40	20	--	80
Workload	40	80	20	80	--

Just how difficult were these three scenarios? Section 4.2 explained that the Combined scenario had five times the activity WU's of the Planning scenario and five times the planning WU's of the Activity scenario. Section 5.2 showed that altitude and airspeed deviations for the Combined scenario were much greater than for the Planning scenario, although comparable to those for the Activity scenario. Earlier in this section, the five subjective ratings for the Combined scenario were all shown to be not only greater than those of the Activity and Planning scenarios, but roughly equal to the sums of the other two.

Was this Planning scenario more "difficult" than that of the preliminary experiments? Altitude deviations were lower by a statistically significant amount, but this can probably be attributed to the use of the autopilot this time. Activity Level and Stress ratings were not statistically different, but Complexity, Difficulty, and Workload were. Difficulty and Workload were greater, at a confidence level of 80 percent. Complexity was greater with 90 percent confidence.

Since both were flown without autopilot, the two Activity scenarios are more easily compared. Altitude deviations were greater this time and both mean absolute and rms differences were greater with 80 percent confidence. Complexity, Stress, and Workload ratings were not statistically different, but Activity Level and Difficulty ratings were. Difficulty ratings were higher for the latest series, at an 80 percent confidence level. Activity Level ratings were also higher, at a 95 percent level.

The only scenario which consistently "saturated" pilots was the Combined scenario. If one defines a "saturated" pilot as one who scores a subjective rating category at 9.0 or higher, the Activity scenario was least likely to saturate pilots. This is interesting because when there were significant differences between the Activity and Planning scenario ratings, the Activity scenario rating was always slightly higher. Thus, certain <u>individuals</u> found the Planning scenario very difficult, while the pilots as a <u>group</u>, found the Planning scenario slightly less demanding than the Activity scenario.

For the Activity scenario, there was one saturated rating for Workload. For the Planning scenario, there were two saturated ratings for Activity Level, and one each for Difficulty and Workload. For the Combined scenario, there were five saturated ratings for Activity Level and Workload, four for Difficulty and Stress, and two for Complexity.

These experiments verified that on a subjective level, a difficult, purely mental task load can equal a difficult, purely manual task load. In general, all the subjective category ratings were similar for the Planning and Activity scenarios.

There was no consistent correlation between subjective ratings and a pilot's experience level. This is not surprising since there is no universal subjective mental metric. Two persons working equally hard may rate their workloads very differently. They have different utilities, and one person may use a linear scale while another uses a logarithmic, and still another, an exponential scale.

Finally, unlike the altitude and airspeed data, all of the subjective rating categories showed monotonically increasing ratings for Segments I, II, and III. This relationship was valid for all scenarios.

5.4 ALTITUDE AND AIRSPEED DEVIATION DATA VERSUS SUBJECTIVE RATINGS

I attempted to correlate altitude or airspeed deviations with each pilot's subjective ratings. However, on an individual basis, objective performance data and subjective ratings were

uncorrelated. This result was not unexpected, and had been reported previously. See, for example, the short discussion in Kantowitz, Hart, and Bortolussi [11]. One possible reason is the "accumulator effect" discussed in Section 1.7. Another reason is that no two people have exactly the same internal metric for rating mental workload. A third reason, suggested in [11], is that one rating may be measuring instantaneous workload while the other is measuring average workload.

One example of the "accumulator effect" can be seen in the data for Pilot A. In flying the Activity scenario, his mean altitude deviations for Segments II and III were 111.8 feet and 106.0 feet: relatively constant. However, his corresponding Workload ratings went from 5.8 for Segment II to 7.1 for Segment III. Thus, his perceived workload was not equal to his performance.

There were also examples of several pilots having similar performance but very different perceived workloads. For instance, in Segment III of the Activity scenario, Pilots B and D had mean altitude deviations of 143.6 feet and 147.8 feet, respectively. However, Pilot B rated his workload at 8.3 while Pilot D rated his only a 6.5.

Nevertheless, in the <u>aggregate</u>, objective performance data <u>was</u> correlated with subjective ratings. Using Pearson's Product-Moment Correlation Coefficient, "r", rms altitude errors weakly correlated with the corresponding subjective ratings for the Activity scenario (See Table 20). Activity Level, Complexity, and Difficulty correlated with an "r" of 0.8 (.805; .797; .807). For the Stress and Workload ratings, "r" was about 0.9 (.911; .903).

Table 20: Pearson Product-Moment Correlation Coefficient for aggregate Altitude Deviations and Subjective Ratings

SCENARIO	Activity		Planning		Combined	
DEVIATION TYPE	mean	rms	mean	rms	mean	rms
Activity Level	.401	.805	.880	.782	.986	.953
Complexity	.389	.797	.843	.777	.999	.896
Difficulty	.403	.807	.817	.746	.990	.945
Stress	.583	.911	.428	.792	.986	.954
Workload	.568	.903	.882	.823	.999	.911

Correlations were slightly better for the Planning scenario. Mean absolute altitude deviations and Activity Level had an "r" of .880. Complexity, Difficulty and Workload had "r's" of .843, .817, and .882. Mean altitude errors did not correlate with Stress, but rms errors did: .792. The ability of the rms error data to correlate with Stress ratings better than the mean deviation data did might be due to the fact that the rms data weights large errors more heavily than small errors. Intuitively, beyond a certain point, stress should be an exponential function of the magnitude of deviations. Thus, large deviations would be better reflected in the rms values and Stress ratings.

There was excellent correlation between mean absolute error data and all five ratings for the Combined scenario. The lowest "r" was for Stress, (.986) with Complexity having an "r" of .9999. Because the pilots were heavily loaded during the Combined scenario, they may have been operating near their personal limits. This may have eliminated the "accumulator effect" and resulted in the good correlation between objective performance data and the subjective ratings.

Finally, Tulga and Sheridan [27] reported that once a subject passed "saturation", performance deteriorated sharply. (Also see Figure 4) While flying the Planning scenario, Pilot C crashed during Segment III. Table 21 lists relevant data for Segments I, II, and III for this pilot. Although he reported only low Stress, the other four subjective factors sharply increased from Segment II to Segment III. Likewise, note that his mean absolute and rms altitude errors increased by 85 percent and 83 percent, and the corresponding airspeed errors increased by 78 percent and 74 percent from Segment II to Segment III. Although one can argue about which was cause and which was effect, mental saturation accompanied a severe performance degradation.

Table 21: Example of related performance deterioration and subjective saturation: Pilot C; Planning Scenario

	SEGMENT		
	I	II	III
Activity Level	5.8	7.4	9.6
Complexity	6.5	6.8	8.3
Difficulty	4.5	4.1	11.0
Stress	1.1	3.0	3.1
Workload	5.5	5.6	10.0
Altitude Error: Mean	11.9	59.8	110.6
RMS	12.7	60.3	110.6
Airspeed Error: Mean	1.2	4.1	7.3
RMS	2.1	4.2	7.3

5.5 PLANNING/MEMORY TASK PERFORMANCE

The manner in which each pilot complied with planning task requests was recorded during each run. This information provided insight into the mental workload problem, and generated some objective data on the mental process.

As workload increased, there were a number of ways that each pilot could respond to these requests. They could fail to perform a task, choosing not to do it or simply forgetting to do it. They could perform the task incorrectly or do some unrequested task. Or, the task might be performed at some time other than the directed time.

Table 22 lists the percentage of planning tasks not performed correctly for each scenario. This data is listed for each pilot and segment. Overall error percentages are plotted in Figure 51.

Although the planning workload for the Baseline and Activity scenarios was the same, the overall error percentage was much higher for the Activity scenario. Similarly, although the Planning and Combined scenarios had similar planning workloads, the Combined scenario percentage was much higher. The Planning and Activity scenarios had similar Subjective ratings, but their mental task performance data was very different.

Table 22: Planning Task error percentages

	Pilot						
	A	B	C	D	E	F	G
BASELINE							
Segment I:	0	0	33	0	0	0	0
II:	NO TASKS						
III:	0	0	0	0	0	0	0
ACTIVITY							
Segment I:	0	100	100	0	0	100	0
II:	NO TASKS						
III:	100	100	100	0	100	100	0
PLANNING							
Segment I:	0	0	0	0	50	25	0
II:	0	29	0	14	0	14	14
III:	7	14	Crash	0	0	18	0
COMBINED							
Segment I:	67	0	33	0	0	67	0
II:	43	71	71	43	43	29	43
III:	64	57	86	29	21	57	7

The Combined scenario results were statistically different from the Planning scenario results at a 99 percent confidence level. Activity scenario results differed from the Combined scenario at an 80 percent confidence level.

As an examination of Table 22 would indicate, the standard deviations for the overall error percentages varied widely from scenario to scenario. For the Baseline and Planning scenarios where the error percentages were low, standard deviations were only 8.8 and 13.4 percent respectively. The difficult Combined scenario had a standard deviation of 27.2 percent, indicating more variability among the pilots. The Activity scenario showed the greatest variability. The low number of mental tasks and the high error percentages for some pilots resulted in a standard deviation of 51.4.

Examining the error data for each segment, the performance for the Planning and Combined scenarios was virtually identical for

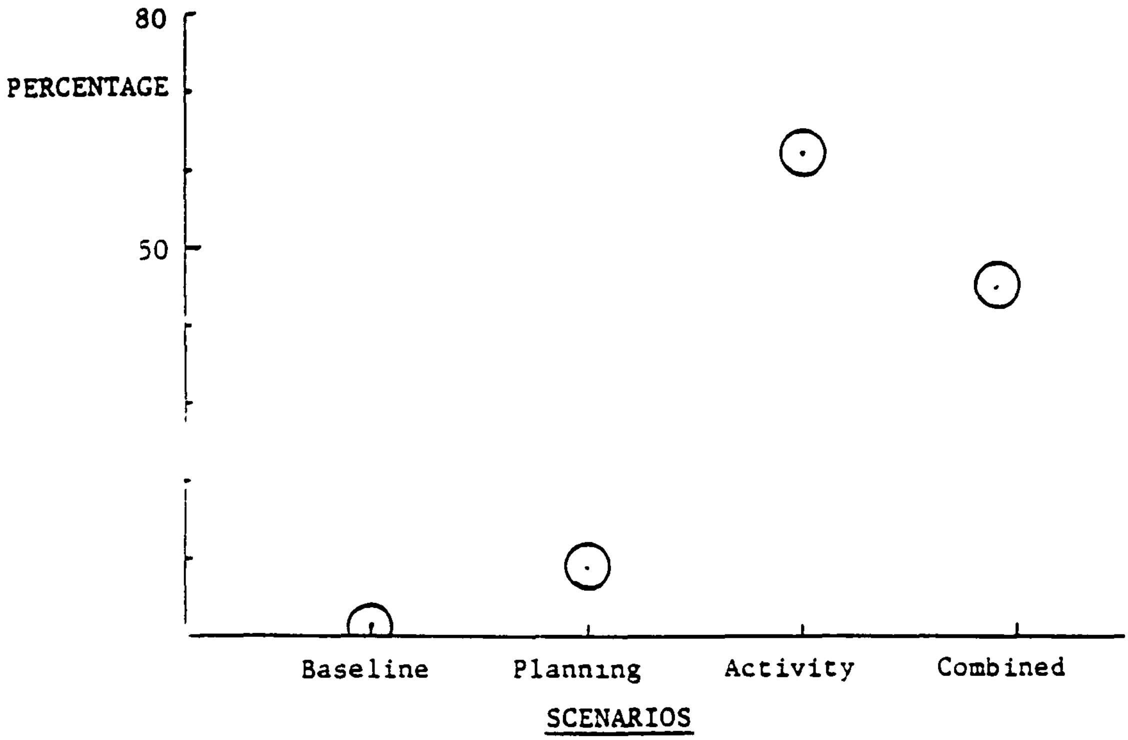


Figure 51: Overall percentage of planning/memory task errors

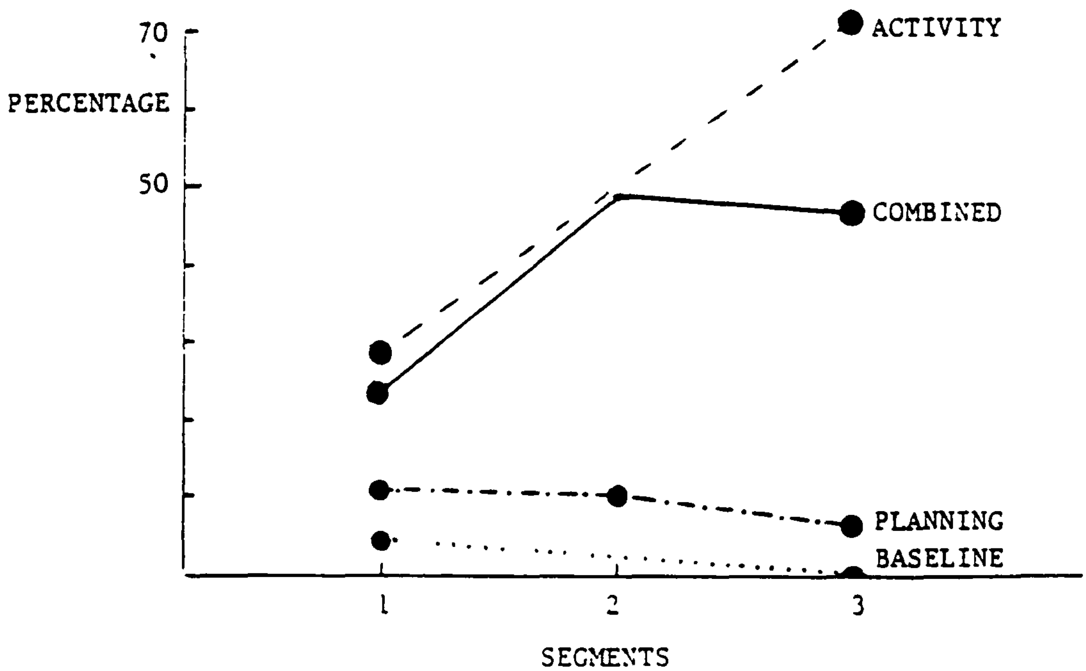


Figure 52: Percentages of planning/memory task errors per segment

Segment I. However, for Segments II and III, the difference between the two scenarios was significant at the 99.9 percent confidence level. On a segment by segment basis, there are too few data points to provide standard deviation data of any value. In general, pilot performance fluctuated a great deal at this level. Figure 52 illustrates the error percentages for each segment and scenario.

The data suggests that at low or moderate levels, manual control workload does not affect mental performance. Sufficient cognitive reserve exists to handle all tasks. However, at relatively high manual control levels, cognitive reserves disappear and mental performance deteriorates. Figure 51 suggests that this mental deterioration may even be evident for low levels of mental tasking.

The preliminary experiments showed a distinct difference in the degree pilots complied with "positional" and "non-positional" memory tasks. A positional task concerned the aircraft's state. For example, the task might be to climb 1000 feet at some point. Non-positional tasks were other types of requests, such as to contact ARTCC at some point.

The preliminary experiment results were unexpected. However, there were not a large number of memory tasks in the preliminary experiments, so the results were statistically suspect. Therefore, this new set of experiments was designed to better illustrate any differences between the two types of tasks by drastically increasing the number of such tasks. Table 23 lists the percentage of each type of task not performed correctly. The data is broken down for each pilot by segment and type of task.

Statistical analysis of the data showed no significant difference in pilot compliance with these two types of tasks. There was, however, a weak indication (70 percent confidence level) that error percentages for both types of tasks increased as workload increased. That is, errors were more likely in Segment III than in Segment I.

This information was also examined to see if the length of time between task request and task execution made any difference for these two types of planning tasks. No statistically significant differences were found.

The various planning tasks were also categorized as Long-term, Medium-term, or Short-term based upon the length of time the pilot had from receiving the task to performing it. Table 24 lists the percentages of improperly performed tasks for these three time periods. The data is only for the Combined and Planning scenarios because the Baseline and Activity scenarios simply had a few Medium-term tasks. Table 24 contains data for each pilot and each segment.

Table 23: Percentage of Positional (P) and Non-Positional (NP) task errors

			Pilot						
			A	B	C	D	E	F	G
PLANNING									
Segment	I:	P	0	0	0	0	50	25	0
		NP	--	--	--	--	--	--	--
	II:	P	0	67	0	33	0	33	0
		NP	0	0	0	0	0	0	33
	III:	P	0	0	Crash	0	0	0	0
		NP	9	18	Crash	0	0	23	0
Overall:		P	0	25	0	13	13	19	0
		NP	7	13	0	0	0	17	7
COMBINED									
Segment	I:	P	67	0	33	0	0	67	0
		NP	--	--	--	--	--	--	--
	II:	P	50	100	75	25	25	25	25
		NP	33	33	67	67	67	33	67
	III:	P	100	80	80	40	60	20	0
		NP	44	44	89	22	11	78	11
Overall:		P	75	67	67	25	33	33	8
		NP	42	42	83	33	25	67	25

Table 24: Error percentages for Long-term (L), Medium-term (M), and Short-term (S) planning tasks

		Pilot						
		A	B	C	D	E	F	G
PLANNING								
Segment I:	S	0	0	0	0	0	50	0
	M	0	0	0	0	100	0	0
	L	--	--	--	--	--	--	--
Segment II:	S	0	0	0	0	0	25	0
	M	0	50	0	50	0	0	50
	L	0	100	0	0	0	0	0
Segment III:	S	11	0	Crash	0	0	28	0
	M	0	33	Crash	0	0	0	0
	L	0	50	Crash	0	0	0	0
Overall:	S	7	0	0	0	0	29	0
	M	0	33	0	17	17	0	17
	L	0	67	0	0	0	0	0
COMBINED								
Segment I:	S	67	0	33	0	0	67	0
	M	--	--	--	--	--	--	--
	L	--	--	--	--	--	--	--
Segment II:	S	75	50	100	25	75	50	50
	M	0	100	50	50	0	0	50
	L	0	100	0	100	0	0	0
Segment III:	S	78	56	89	44	33	33	0
	M	33	67	67	0	0	100	33
	L	50	50	100	0	50	100	0
Overall:	S	81	44	81	31	38	44	13
	M	20	80	60	20	0	60	40
	L	33	67	67	33	33	67	0

When aggregated for each scenario, this data yields the plot shown in Figure 53. Analyzing the error percentages, there was no statistically significant difference within each scenario for the three different task time spans. This was probably because the pilots were allowed to take notes. Additional errors probably arose in the Short-term tasks when the pilots struggled to plan and perform these tasks in a very busy environment. Thus, they would miss some tasks or perform them late. This balanced the errors engendered in the Long-term tasks by the pilots forgetting about tasks.

This hypothesis is supported by the data in Table 25. It lists the number of errors committed by each pilot for each task time span. These errors are classified as errors of Omission (O: did nothing), Commission (C: did something wrong), or Timing (T: did something too early or too late). Note that a large number of the short-term and medium-term errors were the result of timing, whereas no long-term errors were due to mistiming.

However, planning task errors for all three time spans were affected by manual-control activity. Note in Figure 53 that the two low manual workload scenarios (Baseline and Planning) had low error percentages while both high manual workload scenarios (Activity and Combined) had high error percentages. The Activity scenario had a high error percentage even though its planning workload was low.

Looking only at the two high planning workload scenarios, (Planning and Combined) the differences bewtween the scenarios was statistically significant for all three time spans. Differences were significant at an 80 percent confidence level for medium-length tasks, at a 95 percent level for long-term tasks, and 98 percent level for short-term tasks. Thus, the level of manual control was again decisive in determining mental performance.

I chose not to plot or list the standard deviations for this segment-by-segment data. Once again, the data was too coarse and individual pilot performance was too variable to make this information useful.

Figures 54, 55, and 56 illustrate Short-term, Medium-term, and Long-term error percentages for each Segment and scenario. Examining Figure 54, differences between the Planning and Combined scenarios for Short-term planning tasks were not statistically significant in Segment I. Differences were at a 70 percent confidence level. However, the differences were at a 98 percent confidence level for Segments II and III, when workloads were higher.

Referring to Figure 55 for Medium-term task results, differences between the Planning and Baseline or between the Planning and Activity scenarios were insignificant for Segment I (20 percent confidence level). The Planning and Combined scenario

Table 25: Types and numbers of Long-term, Medium-term, and Short-term planning errors

SCENARIO		Pilot A	B	C	D	E	F	G
PLANNING								
Long:	O:	.	1	.	.	.	.	.
	C:	.	1	.	.	.	.	.
	T:	.	.	.	.	.	.	.
Medium:	O:	.	.	.	.	.	.	.
	C:	.	1	.	1	.	.	.
	T:	.	1	.	.	1	.	1
Short:	O:	.	.	.	.	.	.	.
	C:	.	.	.	.	.	1	.
	T:	1	.	.	.	.	4	.
Overall:	O:	.	1	.	.	.	.	.
	C:	.	2	.	1	.	1	.
	T:	1	1	.	.	1	4	1
COMBINED								
Long:	O:	1	2	2	1	1	2	.
	C:	.	.	.	.	.	.	.
	T:	.	.	.	.	.	.	.
Medium:	O:	.	2	3	.	.	1	1
	C:	.	.	.	.	.	.	.
	T:	1	2	.	1	.	2	1
Short:	O:	8	4	12	4	5	4	.
	C:	.	.	.	.	.	.	.
	T:	5	3	.	1	1	2	2
Overall:	O:	9	8	17	5	6	7	1
	C:	.	.	.	.	.	.	.
	T:	6	5	.	2	1	4	3

Note: O = Omission; C = Comission; T = Timing Error

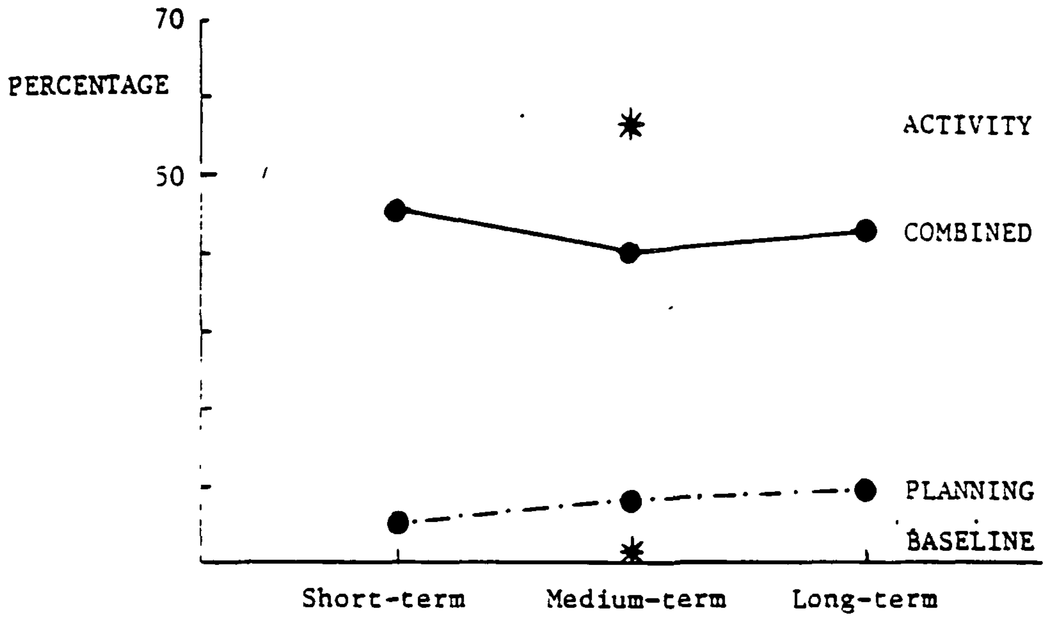


Figure 53: Error percentages for three different planning task time spans

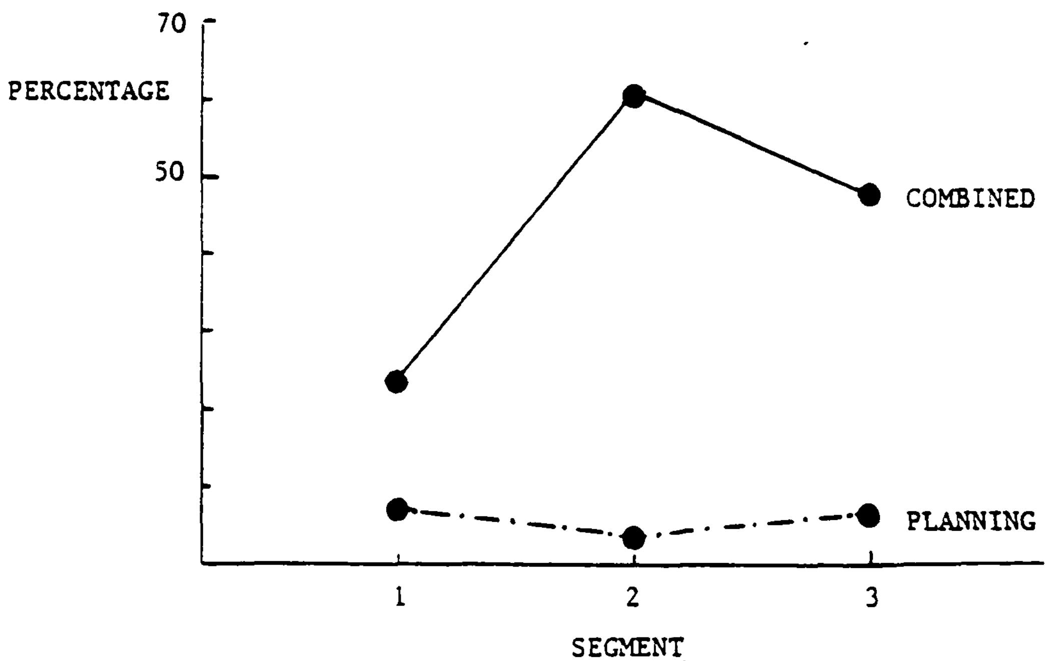


Figure 54: Short-term planning/memory task error percentages by segment

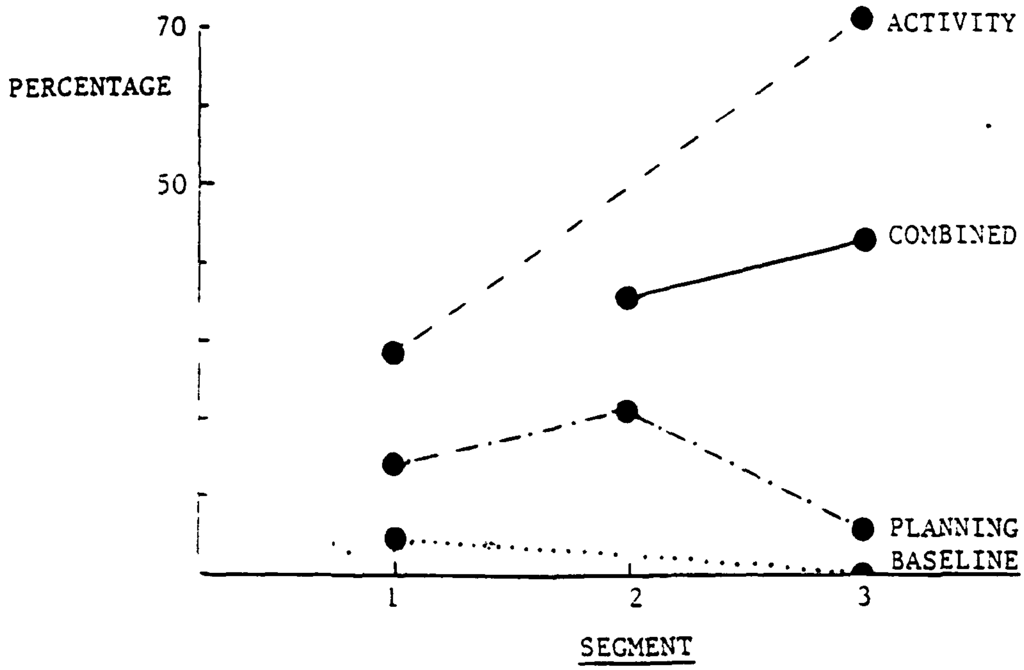


Figure 55: Medium-term planning/memory task error percentages by segment

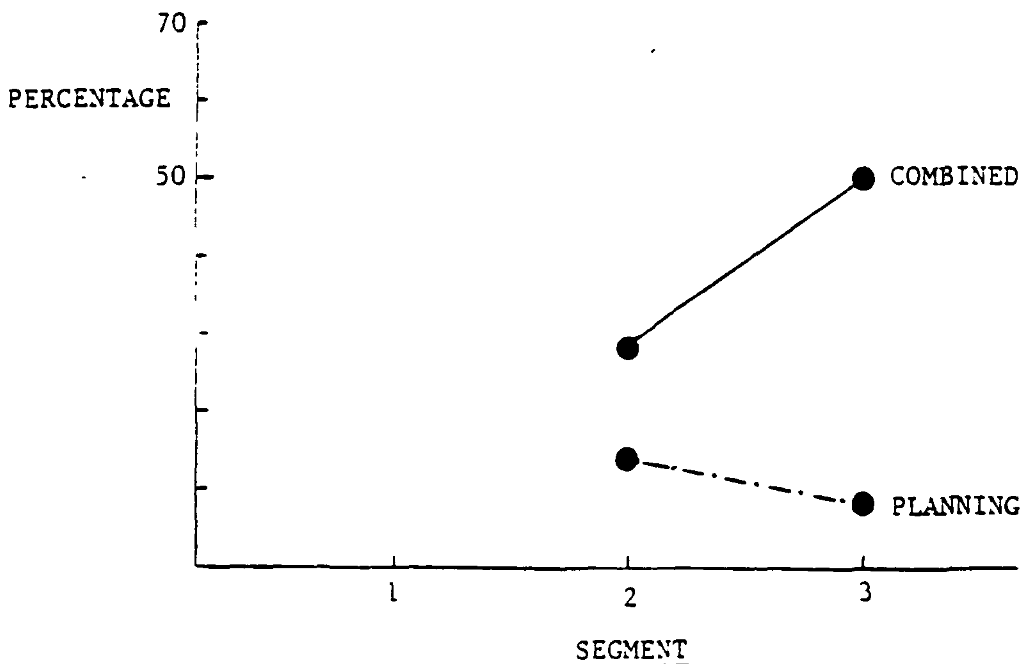


Figure 56: Long-term planning/memory task error percentages by segment

errors were statistically indestinguishable for Segment II. However, in Segment III, the highest workload segment, the Combined scenario errors were higher than the Planning scenario errors (90 percent confidence level). The Planning and Activity difference was even greater: a 95 percent confidence level. The Combined and Activity scenarios were different, but at a much lower confidence level (80 percent). Again, at high overall workload levels, the presence of a high manual workload made a significant difference.

Figure 56 is a plot of the Long-term planning task results. In Segment II, the Planning and Combined scenarios were statistically indestinguishable. However, at the higher workload level of Segment III, the error percentage for the Combined scenario was clearly greater (90 percent confidence level).

The Activity and Planning scenarios had moderate manual or mental workloads. At these levels, error percentages were similar for all of the pilots. However, some differences arose for the high workload Combined scenario. Refering to Table 25, the average number of planning task errors for the low experience and high experience pilots were very different. The low experience pilots (A and B) averaged 14.0 task errors while the high experience pilots (D, E, F, and G) averaged 7.3 task errors. Thus, there were signs of experience related saturation in this mental performance data which was much less obvious in the objective performance data and subjective rating data. This difference was verified at a 95 percent confidence level.

The number of individual planning errors and individual altitude or airspeed deviations were not correlated. Nor were planning errors and subjective ratings. However, in aggregate, all three measures increased with increasing workload.

5.6 PILOT COMMENTS

The planning task instructions given to the pilots were not always in chronological order. This was done to make the planning function more difficult, more complex, and to increase overall workload without further increasing the number of assigned tasks. This strategy apparently worked, since several subjects mentioned that instructions "mixed in time" were difficult to organize.

With the exception of ETA's (Estimated Times of Arrival) and Clearance times, most ATC to pilot instructions are linked to a geographic point. For example, ATC may direct, "climb 1000 feet, now", or "climb to Flight Level 290 at Knoxville VOR". In this experiment, the pilots were usually told to do something at a certain elapsed time. This was done primarily to make the pacing more uniform across runs and subjects, and to aid in analyzing the

data. However, the pilots remarked that tasks required at a certain time were harder to remember than tasks required at a place. This was consistent with their experience and indicated that there was an unintentional, but experimentally welcome, boost in mental workload due to this request format.

Individual pilots found the autopilot to be either a hindrance or a great help. Several pilots stated that when things "really got busy", the autopilot was the only thing which kept workload at a manageable level. But, several pilots reported that having to plan how to use the autopilot was worse than the demanding manual control work. With purely manual control, they said you "simply do what you need to do." It should be mentioned, however, that the pilots who disliked the autopilot had thousands of hours of flight time but little previous experience with autopilots. This appears to be an example of highly skilled operators preferring to function in a familiar mode rather than sit, think, and program a machine: the kind of mental workload problem which initially instigated studies of this type.

There was a general consensus that Mental Workload was best reflected in the Stress and Workload subjective ratings.

A number of the pilots stated that planning and memory items tended to get second priority to immediate task demands. This is consistent with the finding that a high activity workload significantly increased planning task errors. Pilots were obeying the prime directive taught every student pilot: "First, fly the aircraft!" These statements and results are also consistent with Tulga and Sheridan's finding that subjects don't plan ahead when they're very busy [27].

Finally, the pilots mentioned four non-experiment-specific items which increased mental stress and workload. One was the "annoyance" factor caused by having too many things to do or by being interrupted before completing a task. This type of problem is common on final approach when the need to fly and/or monitor equipment, clear for other aircraft, look for the runway, interact with ATC, and run aircraft checklists, combine to make the flight deck a busy, stressful environment.

A second item was the effect of "getting behind". Again, this is most likely to occur when things get very busy. The stress generated by a lengthening "mental queue", combined with the possible need to modify a former plan, increases the perceived workload.

Similarly, abnormal events significantly increase workload, disrupt concentration, and increase the frustration level. These effects have been discussed in the open literature. See, for example, [5], [9], and [25].

The fourth item concerned the effect of adding an increment of workload when the workload is already high. As the pilot becomes task saturated, additional tasks must be prioritized, added to a mental queue, or ignored. This increases stress, frustrates the pilot, and increases his mental manipulations. These factors result in lower performance, increased mental workload, and lower safety margins. For additional discussion, see Tulga and Sheridan [27].

Chapter 6

FINDINGS AND RECOMMENDATIONS

6.1 MAJOR FINDINGS

1. The number of additional assigned mental tasks had no statistically significant impact on the degree of aircraft control. The level of manual workload was the decisive factor. When mental workload was high but manual workload was low, altitude and airspeed deviations were small. When mental workload was low but manual workload was high, altitude and airspeed deviations were large.

2. Incremental subjective ratings were calculated relative to the ratings for a Baseline scenario. The incremental rating for a high manual workload scenario added to the incremental rating for a high mental workload scenario was equal to the incremental rating for a scenario which combined both types of workloads.

3. The type of scenario (manual or mental) and the degree of workload determined whether the five Subjective Rating categories (Activity Level, Complexity, Difficulty, Stress, and Workload) were perceived as similar or different. The pilots found differences in the meanings of the five categories for a scenario with a moderately high <u>manual</u> workload. When <u>mental</u> workload was moderately high, Stress ratings were similar to the Difficulty ratings, and Workload ratings were similar to the Complexity ratings. For a combination of very high <u>manual</u> <u>and</u> <u>mental</u> workloads, Activity Level and Stress were distinguishable, but distinctions among the other ratings were blurred. Workload and Difficulty were correlated, and Stress and Difficulty ratings were similar.

4. Subjective ratings given by individual pilots during the high manual workload scenario were very similar. However, there were individual differences in the subjective ratings for the high mental workload scenario. Some pilots were not stressed by the mental tasks while others significantly increased their subjective ratings.

5. At low or moderate manual and mental workload levels, aircraft deviations and memory task performance did not correlate with the subjective ratings. At high workload levels, the correlation was very good. It's possible that at lower workloads, there is reserve mental capacity which varies from pilot to pilot, affecting performance and ratings. At high workload levels, all pilots may be tapping most or all of their mental capacity, resulting in much greater consistency between performance and the

subjective ratings.

6. The magnitude of manual workload was decisive in determining the ability of the pilots to handle mental tasks. A mentally difficult, manually easy scenario resulted in a low percentage of mental errors. A mentally easy, manually difficult scenario resulted in a high percentage of mental errors. The manual activity was presumably consuming a great deal of the pilots' mental processing capacity, even when they were not aware of it. This finding was equally valid for long-term, medium-term, and short-term mental tasks.

7. Under conditions of high manual and mental workload, the low experience pilots did not perform mental tasks as well as the high experience pilots did. However, objective performance and subjective ratings were similar for the two groups. Thus, these experiments suggest that monitoring and measuring mental performance might be a more sensitive indicator of mental workload and reserve mental capacity than the other measures.

6.2 RECOMMENDATIONS

These experiments produced a mountain of raw data. I analyzed a great deal of the data and examined the relationships between many different variables. However, I did not exhaust all possibilities. There are still a number of variables which could be compared, examining correlations and differences.

It may be enlightening to "filter" the airspeed and altitude data. I measured <u>all</u> deviations to derive mean absolute errors and rms errors. Although, in theory, all pilots strive to maintain desired altitudes and airspeeds perfectly, they often induce small errors to provide sensory feedback and gain additional information on the aircraft's performance. In addition, pilots tend to fly within individual tolerances. These tolerances change, depending on such factors as height above the ground, airspeed stall margin, meteorological conditions, physical and mental states, and a number of personal factors which affect an individual's utilities.

One might filter the altitude and airspeed data to account for these tolerances. Considering all airspeed deviations less than $\pm$ 5 knots and all altitude deviations less than $\pm$ 50 feet as zero deviations may radically change the results, better separating the low from the high experience pilots, or more readily determining which pilots were saturated. For actual flight checks, permissable performance is usually $\pm$ 10 knots and $\pm$ 100 feet. Using these limits would provide a still coarser data set. Comparisons between results obtained from such data and this study might be enlightening.

Subjective Ratings should be used in future studies of mental workload. They provide a useful, if imprecise, measure of the pilot's mental state.

The only significant difference found between the low experience and high experience pilots was in their performance of mental planning tasks. This should be further investigated in future studies.

This study also found that objective manual performance data and subjective ratings were correlated at high workload levels. If verified by a new series of experiments, this might provide a useful group metric for mental workload.

There was a linear relationship between the subjective ratings of scenarios with different workloads when those ratings were measured relative to a baseline scenario. The ratings for a scenario with high manual workload were added to the ratings of a mentally difficult scenario and found equal to the ratings of a scenario which combined those manual and mental tasks. Future studies should test and define the limits of this apparent linearity.

For future variations on these experiments, several changes may be useful. First, the experimenter may choose to add an aircraft checklist. When the pilots were approaching the Localizer course or on final approach, they had to fly, make radio calls to a simulated ATC, and configure the aircraft, navigational aids, and autopilot. However, they did not have a checklist to process. When added to all the other necessary tasks, this "necessary evil" can be a significant burden on final approach. Adding such a checklist would also increase the realism of the simulated flight environment.

Second, further examinations of the effect of memory time span on mental workload should eliminate the medium-term tasks and concentrate on short-term versus long-term differences. It might also be beneficial to use fewer simultaneous mental tasks and to eliminate the note pad which I provided for the pilots' use. This would further emphasize the <u>memory</u> aspect of mental workload while I emphasized the <u>planning</u> component of mental workload.

Third, I recommend eliminating the autopilot from future experiments. Not only will this make it easier and less time consuming to train future volunteers, but it will also reduce one variable, simplifying analysis. Furthermore, based on the results of these experiments, eliminating its use would help keep the manual workload level high. The most interesting effects were found at high manual workload levels.

Anyone analyzing this data for a future study might consider eliminating the data for Pilot C. Although he was near the mean in terms of experience, he consistently had the greatest altitude and

airspeed deviations and the largest number of mental task errors. In addition, his subjective ratings were consistently at the extremes of the group's ratings. In fact, his ratings were usually abnormally low, indicating that he thought the scenarios were easier than did the other pilots.

Finally, a researcher might perform multivariate analyses and employ other sophisticated mathematical techniques to further examine performance, perceptions, and the interrelationship of the two.

Appendix 1

SIMULATION AND FLIGHT DYNAMICS

A-1.1 SIMULATION FLIGHT DYNAMICS

The basic flight dynamics for the aircraft simulation were modelled on the Lockheed Jetstar, a four engine business jet. The Jetstar's longitudinal and lateral stability derivatives were obtained from NASA CR-2144 [6]. The simulation used the coefficients for Mach 0.230 (152 kts) at Sea Level. This provided good fidelity for final approach flight characteristics and did not adversely affect handling qualities until airspeed exceeded Mach 0.340 (225 kts). Beyond Mach 0.340, handling gradually becomes more sensitive.

Figure 57, taken from McRuer, Ashkenas, and Graham [10], shows the nomenclature used for defining the stability derivatives' velocities, forces, and moments. Table 26 gives the desired longitudinal and lateral stability derivatives. The derivatives in the NASA document were in English units (feet, radians, seconds) and in body axes. These coefficients were translated into stability axes and then converted into MKS units (meters, kilograms, seconds) for the simulation.

The linearized differential equations in Laplace form for the flight dynamics are:

Longitudinal Dynamics:

$$\begin{bmatrix} s-X_u & -X_{\dot\alpha}*s-X_\alpha & -X_q*s+g*\cos(gam_0) \\ -Z_u & (U_0-Z_{\dot\alpha})s-Z_\alpha & (-U_0-Z_q)s+g*\sin(gam_0) \\ -M_u & -M_{\dot\alpha}*s-M_\alpha & s(s-M_q) \end{bmatrix} \begin{bmatrix} u \\ \alpha \\ \theta \end{bmatrix} = \begin{bmatrix} X_\delta \\ Z_\delta \\ M_\delta \end{bmatrix} *\delta$$

Where: α = alpha (angle of attack) [radians]

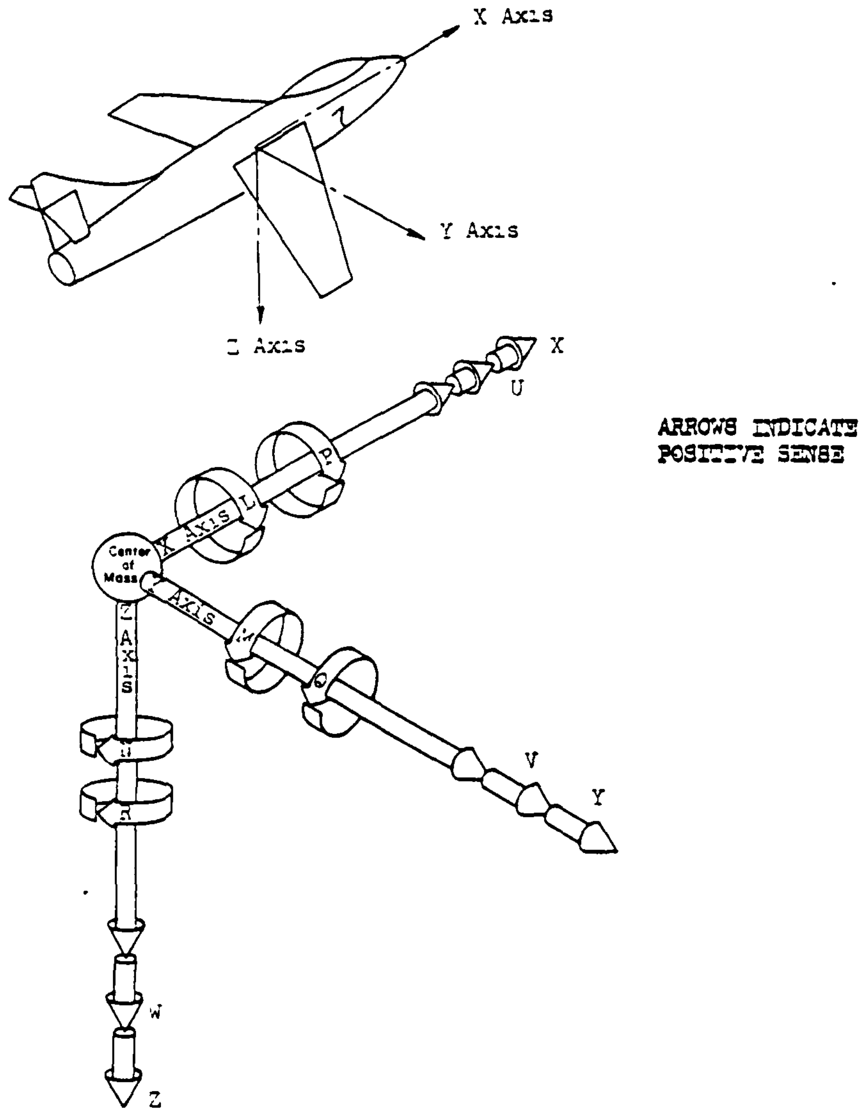
θ = theta (pitch angle) [radians]

u = perturbed airspeed [m/sec] $U = U_0 + u$

g = 9.8 [m/sec^2]

gam_0 = $gamma_0$ (trimmed flight path angle; negative is downward) [radians]

q = pitch angle rate



	Velocities	Applied forces and moments	Distances
Forward	U	X	x
Side	V	Y	y
Vertical	W	Z	z
Roll	P	L	
Pitch	Q	M	
Yaw	R	N	

Figure 57: Vehicle-fixed axis system and notation

Table 26: Aerodynamic Coefficients for the Lockheed Jetstar

U_0 = 78.196 [m/sec] = 152 [knots]

	CR-2144 (body axes)	(stab. axes)	MKS Units
X_u	-0.00456 [1/sec]	-0.02004	X_u = -0.0200417 [1/sec]
X_w	0.164 [1/sec]	0.024815	X_α = 1.94043 [m/sec^2]
$X_{\dot{w}}$	0.0		$X_{\dot{\alpha}}$ = 0.0
X_q	0.0		X_q = 0.0
$X_{\delta_{el}}$	2.78 [ft/sec^{2}rad]	0.0077789	$X_{\delta_{el}}$ = 0.002371 [m/sec^{2}rad]
$X_{\delta_{th}}$	0.000842 [ft/sec^{2}rad]	0.0008259	$X_{\delta_{th}}$ =0.0002517[m/sec^{2}rad]
Z_u	-0.103 [1/sec]	-0.2421889	Z_u = -0.2421889 [1/sec]
Z_w	-0.723 [1/sec]	-0.7075117	Z_w = -55.324584 [m/sec^2]
$Z_{\dot{w}}$	0.0		$Z_{\dot{w}}$ = 0.0 [m/sec]
Z_q	0.0 [ft/radsec]		Z_q = 0.0 [m/radsec]
$Z_{\delta_{el}}$	-14.0 [ft/sec^{2}rad]	-14.27334	$Z_{\delta_{el}}$ = -4.350514[m/sec^{2}rad]
$Z_{\delta_{th}}$	0.0 [ft/sec^{2}rad]	-0.0001604	$Z_{\delta_{th}}$ =-0.0000488[m/sec^{2}rad]
M_u	0.00175 [1/ftsec]	-0.0000353	M_u = -0.0001158 [1/msec]
M_w	-0.00902 [1/ftsec]	-0.0091881	M_α = -2.3613417 [1/sec^2]
$M_{\dot{w}}$	-0.000834 [1/ft]	-0.0008181	$M_{\dot{\alpha}}$ = -0.2102517 [1/sec]
M_q	-0.582 [1/sec]	-0.582	M_q = -0.582 [1/sec]
$M_{\delta_{el}}$	-2.80 [1/sec^{2}rad]	-2.80	$M_{\delta_{el}}$ = -2.80 [1/sec^{2}rad]
$M_{\delta_{th}}$	-0.00000604[1/sec^{2}rad]	-0.00000604	$M_{\delta_{th}}$ =-0.00000604[1/sec^{2}rad]

Table 26, continued

	CR-2144 (body axes)	(stab. axes)	MKS Units
Y_{Beta}	-25.8 [ft/sec^2]	-25.8	Y_{Beta} = -7.86384 [m/sec^2]
$Y_{\dot{B}eta}$	0.0 [ft/sec]		$Y_{\dot{B}eta}$ = 0.0 [m/sec]
Y_p	0.0 [1/rad]		Y_p = 0.0 [1/rad]
Y_r	0.0 [1/rad]		Y_r = 0.0 [1/rad]
$Y_{\delta_{ail}}$	0.0 [ft/sec^{2}rad]		$Y_{\delta_{ail}}$ =0.0 [m/sec^{2}rad]
$Y_{\delta_{rud}}$	0.0244 [ft/sec^{2}rad]	0.0244	$Y_{\delta_{rud}}$ =1.9079824 [m/sec^{2}rad]
L'_{Beta}	-3.42 [1/sec^2]	-3.14121	L'_{Beta} = -3.14121 [1/sec^2]
$L'_{\dot{B}eta}$	0.0 [1/sec]		$L'_{\dot{B}eta}$ = 0.0 [1/sec]
L'_p	-0.752 [1/sec]	-0.7184954	L'_p = -0.7184954 [1/sec]
L'_r	0.234 [1/sec]	0.3422086	L'_r = 0.3422086 [1/sec]
$L'_{\delta_{ail}}$	1.04 [1/sec^{2}rad]	1.0034	$L'_{\delta_{ail}}$ = 1.0034 [1/sec^{2}rad]
$L'_{\delta_{rud}}$	0.533 [1/sec^{2}rad]	0.4101933	$L'_{\delta_{rud}}$ =0.4101933 [1/sec^{2}rad]
N'_{Beta}	1.10 [1/sec^2]	1.74333	N'_{Beta} = 1.74333 [1/sec^2]
$N'_{\dot{B}eta}$	0.0 [1/sec]		$N'_{\dot{B}eta}$ = 0.0 [1/sec]
N'_p	-0.173 [1/sec]	-0.0647908	N'_p = -0.0649708 [1/sec]
N'_r	-0.172 [1/sec]	-0.205504	N'_r = -0.205504[1/sec]
$N'_{\delta_{ail}}$	-0.0864 [1/sec^{2}rad]	-0.2867578	$N'_{\delta_{ail}}$ =-0.2867578[1/sec^{2}rad]
$N'_{\delta_{rud}}$	-0.580 [1/sec^{2}rad]	-0.672237	$N'_{\delta_{rud}}$ =-0.672237 [1/sec^{2}rad]

Lateral Dynamics:

$$\begin{bmatrix} (U_0 - Y_{\dot{Beta}})s - Y_{Beta} & -Y_p * s - g * \cos(gam_0) & U_0 - (Y_r - Y_p * \tan(gam_0)) \\ -L'_{\dot{Beta}} * s - L'_{Beta} & s(s - L'_p) - \tan(gam_0) & s - (L'_r - L'_p * \tan(gam_0)) \\ -N'_{\dot{Beta}} * s - N'_{Beta} & -N'_p * s & s - (N'_r - N'_p * \tan(gam_0)) \end{bmatrix} \begin{bmatrix} Beta \\ Phi \\ r \end{bmatrix} = \begin{bmatrix} Y_{\delta} \\ L'_{\delta} \\ N'_{\delta} \end{bmatrix} * \delta$$

Where: Beta = side slip angle [radians]
Phi = roll angle [radians]
r = yaw rate [rad/sec]
p = roll rate [rad/sec]

$$L'_i = \frac{L_i + (I_{xz}/I_x)N_i}{1 - (I_{xz}^2/I_x I_z)}$$

$$N'_i = \frac{N_i + (I_{xz}/I_z)L_i}{1 - (I_{xz}^2/I_x I_z)}$$

The coefficients from Table 26 were used to generate aircraft dynamics equations in a linear state-variable form: $\dot{x} = Ax + By$. A separate program generated the A and B matrix coefficients for the longitudinal and lateral modes. The matrices for the longitudinal case are:

$$x = [u \ \alpha \ q \ \theta]^T, \qquad y = [\delta_{el} \ \delta_{th}]^T$$

$$A = \begin{bmatrix} X_u + X_{\dot{\alpha}} Z_u / c & X_{\alpha} + X_{\dot{\alpha}} Z_{\alpha} / c & X_q X_{\dot{\alpha}} (U_0 + Z_q)/c & -g * c_0 - X_{\dot{\alpha}} * g * s_0 / c \\ Z_u / c & Z_{\alpha} / c & (U_0 + Z_q)/c & -g * s_0 / c \\ M_u + M_{\dot{\alpha}} Z_u / c & M_{\alpha} + M_{\dot{\alpha}} Z_{\alpha} / c & M_q + M_{\dot{\alpha}} (U_0 + Z_q)/c & -M_{\dot{\alpha}} * g * s_0 / c \\ 0 & 0 & 1 & 0 \end{bmatrix}$$

$$B = \begin{bmatrix} X_{\delta_{el}} + X_{\dot{\alpha}} Z_{\delta_{el}} / c & X_{\delta_{th}} + X_{\dot{\alpha}} Z_{\delta_{th}} \\ Z_{\delta_{el}} / c & Z_{\delta_{th}} / c \\ M_{\delta_{el}} + M_{\dot{\alpha}} Z_{\delta_{el}} / c & M_{\delta_{th}} + M_{\dot{\alpha}} Z_{\delta_{th}} \\ 0 & 0 \end{bmatrix}$$

Where: $c_0 = \cos(\text{gamma}_0)$

$s_0 = \sin(\text{gamma}_0)$

$c = U_0 - Z_{\dot{\alpha}}$

δ_{el} = elevator deflection

δ_{th} = throttle position

For the lateral case, the matrices are:

$$x = [\text{Beta}\ p\ \text{Phi}\ r]^T, \qquad y = [\delta_{ail}\ \delta_{rud}]^T$$

$$A = \begin{bmatrix} Y_{Beta}/h & Y_p/h & g*c_0/h & -(U_0 - Y_r)/h \\ L'_{Beta} + L'_{\dot{B}eta} Y_{Beta}/h & L'_p + L'_{\dot{B}eta} Y_p/h & L'_{\dot{B}eta} * g * c_0/h & L'_r - L'_{\dot{B}eta}(U_0 - Y_r)/h \\ 0 & 1 & 0 & \tan(\text{gamma}_0) \\ N'_{Beta} + N'_{\dot{B}eta} Y_{Beta}/h & N'_p + N'_{\dot{B}eta} Y_p/h & N'_{\dot{B}eta} * g * c_0/h & N'_r - N'_{\dot{B}eta}(U_0 - Y_r)/h \end{bmatrix}$$

$$B = \begin{bmatrix} Y_{\delta_{ail}}/h & Y_{\delta_{rud}}/h \\ L'_{\delta_{ail}} + L'_{\dot{Beta}} Y_{\delta_{ail}}/h & L'_{\delta_{rud}} + L'_{\dot{Beta}} Y_{\delta_{rud}}/n \\ 0 & 0 \\ N'_{\delta_{ail}} + N'_{\dot{Beta}} Y_{\delta_{ail}}/h & N'_{\delta_{rud}} + N'_{\dot{Beta}} Y_{\delta_{rud}}/h \end{bmatrix}$$

Where: $h = U_0 - Y_{\dot{Beta}}$

δ_{ail} = aileron deflection

δ_{rud} = rudder deflection

A-1.2 SIMULATOR SOFTWARE

Figure 58 is a basic diagram of the data sets and subroutines for the simulator software.

The program begins by calling INITL. INITL sets certain constants, initializes the input keyboard, and configures the MEGATEK display. To do this, INITL begins by calling MSETUP, a subroutine for initializing and readying the MEGATEK for display input. INITL next reads PERSV.DAT, a data set which contains information needed by the MEGATEK for drawing a perspective display. Then, PERSIN initializes the display with the PERSV data. KINIT initializes the inputs from the experimenter's keyboard. INITL then reads BDAT.DAT. This data set contains constants which initialize the aircraft's position, state, and configuration. It also sets navigational aid coordinates and input parameters for the aircraft Control Box. COEF then reads the aircraft's longitudinal and lateral dynamics coefficients from MATR.DAT.

The main program next calls ATCIN. ATCIN reads ATCN.DAT which sets wind speeds, wind directions, wind regions, ceiling height, and initializes certain automatic Air Traffic Control situations and capabilities. These capabilities were not used in these experiments.

Once everything is initialized, the main program calls the FLY subroutine: the actual flight simulation. FLY calls INPUT, a subroutine which allows the experimenter to interrupt the simulation, change conditions, or begin or terminate data storage. INPUT calls DATREC which decides whether or not to store data. If the inputs or conditions are proper for storing data, DATREC calls

DREC which records the data in memory. INPUT then calls KEY, a subroutine which looks for and accepts keyboard inputs.

FLY then calls AUPI which determines the autopilot configuration and autopilot pitch, roll, and throttle commands. For more information on autopilot functions and dynamics, see Appendix 2.

Next comes the CWS routine. CWS stands for Control Wheel Steering. When CWS is activated, it acts as a stability augmentation system, smoothing aircraft dynamics. For more information, see Appendix 2.

DYNAM1 calculates vehicle longitudinal response based on autopilot or control box inputs, aircraft state, aircraft configuration, and the linear state-variable matrices obtained from MATR.DAT. DYNAM2 performs a similar calculation for the vehicle's lateral response.

The NAVIGA subroutine takes the data on the change in aircraft state and uses it to calculate the new position, state, and position relative to the VOR/DME or ILS selected on the Control Box.

OUTPUT updates the MEGATEK display and stores data. OUTPUT first calls DISPLY. DISPLY calls PERS, which updates the perspective display. Then DISPLY performs calculations on all the relevant data to update the flight instrument display on the MEGATEK. If requested, OUTPUT will store the desired data in a new data set.

Although it wasn't used for these experiments, the program contains a major subroutine called ATC. ATC provides a capability for generating automatic ARTCC instructions on the MEGATEK display. This capability was not used in these experiments because the high pilot workload would have made it difficult for the pilots to read the instructions. Also, since there were no audio cues available to alert the pilot when an instruction appeared on the screen, the chance of a busy pilot missing an instruction was great. It would have been extremely difficult to determine if the pilots failed to do something because they forgot to or if they simply missed the instruction. This would have been unsatisfactory since a major part of these experiments involved measuring the frequency with which pilots forgot instructions.

ATC first calls AIRP which provides the current aircraft position. Then, RPR determines which ARTCC sector the aircraft is in. Sectors are defined in ATCN.DAT. RPR calls GCL to determine if any ground controller instructions should be issued. If instructions are necessary, GCL calls ATCR which generates the desired ARTCC directions.

At this point, the main program returns to the FLY subroutine

and continues looping until a stop command is issued through the keyboard and registered by FLY's INPUT subroutine.

```
MAIN
    INITL
        MSETUP
        Read PERSV.DAT
        PERSIN
        KINIT
        Read BDAT.DAT
        COEF
            Read MATR.DAT
    ---------------------------
    ATCIN
        Read ATCN.DAT
    ---------------------------
    FLY
        INPUT
            DATREC
                DREC
            KEY
        AUPI
        CWS
        DYNAM1
        DYNAM2
        NAVIGA
        OUTPUT
            DISPLY
                PERS
    ---------------------------
    ATC
        AIRP
        RPR
            GCL
            ATCR
```

Figure 58: Principle simulation routines, subroutines, and data sets

Appendix 2

AUTOPILOT AND STABILITY AUGMENTATION SYSTEMS

A-2.1 CWS (CONTROL WHEEL STEERING) DYNAMICS

The CWS system is an optional flight control mode which provides an inner feedback loop to improve aircraft control characteristics. In Laplace form, the elevator and throttle commands are generated by the following relations:

$$\delta_{el} = 3.0q + 4.0\ \theta + \frac{6.88\ \delta_{col}}{1.0 + 0.4s}$$

$$\delta_{th} = 5000.0\ \frac{\delta_{ct}}{1.0 + 1.0s}$$

Where: δ_{col} = pitch command

δ_{th} = throttle command

In the elevator command equation, q (pitch rate) feedback improves stability. The stick pitch command term makes the elevator command proportional to stick position. The first-order lag term in the throttle command equation simulates engine response lag.

The CWS aileron command for roll control is:

$$\delta_{ail} = 4.5p + 0.39375r + 6.75\text{Phi} + \frac{5.0625s\ \delta_w}{1.0 + 5.0s} + \frac{2.53125\ \delta_w}{s}$$

Where: δ_w = lateral stick command

There are roll rate, yaw rate, and roll angle feedback terms in the aileron command. The last two terms are important because they result in the bank angle being proportional to the integrated value

of the stick deflection.

Finally, the CWS rudder command is:

$$\delta_{rud} = -3.0 \text{ Beta} - 2.3 \dot{\text{Beta}}$$

The rudder command is a function of yaw angle and yaw rate.

The simple "mechanical" ratios for the system are:

$$\delta_{el} = 6.88\ \delta_{col}$$

$$\delta_{th} = 5000.0\ \delta_{ct}$$

$$\delta_{ail} = 2.25\ \delta_{w}$$

A-2.2 AUTOPILOT DYNAMICS

A-2.2.1 MANUAL HEADING MODE

The lateral autopilot's manual heading mode allows the pilot to command a magnetic heading by turning a knob on the Control Box. Turning the control knob slews an indicator on the HSI. The autopilot will turn the aircraft in the shortest direction to the heading set under the indicator. The stick deflection command signal is:

$$\delta_w = 2.5 \left[\left[1.2 \text{ (psic - psi)} \right]_{0.4} - \text{Phi} \right]$$

Where: psic = commanded heading (radians)

psi = present magnetic heading (radians)

This mode will roll the aircraft into a 23° bank angle for any heading error greater than approximately 19°. For errors less than 19°, bank angle is proportional to heading error. Note: when this mode is engaged, any pilot lateral stick inputs are ignored.

A-2.2.2 VOR COUPLE MODE

In this lateral autopilot mode, the aircraft will turn to intercept a chosen magnetic course relative to a selected VOR. The Laplace equation which determines the stick deflection command is:

$$\delta_w = 2.5\left[\left[\left[\frac{(-0.001)\mathrm{DME}(1.0 + 23.0s)}{(1.0 + 1.0s)\ \mathrm{VORE}} - 1.0\ \mathrm{VCRSE}\right]_{.75} - 1.0\ \mathrm{VCRSE}\right]_{.4} - \mathrm{Phi}\right]$$

Where: DME = distance from the VOR or runway

VORE = difference between the selected VOR course radial and the current VOR radial (radians)

VCRSE = angular difference between the selected VOR course and the current aircraft magnetic heading (radians)

The innermost bracket acts on the rate of change between the desired VOR radial and the present one. This rate is artificially limited to reduce sensitivity near the VOR. However, since this bracket acts like a differentiator, it still produces rapid corrections. The next bracket outward serves to limit the bank angle response to a maximum of 23°. The outermost bracket commands a stick deflection signal proportional to the difference between the desired bank angle and the actual bank angle. The overall effect is to command the bank angle to a certain value which is proportional to the rate of VOR error. Pilot lateral stick inputs are ignored in this mode.

A-2.2.3 LOCALIZER COUPLE MODE

The Localizer couple mode functions identically to the VOR couple mode. The only difference is that errors are measured relative to a Localizer course instead of a VOR course. All dynamics and limits are identical.

A-2.2.4 ALTITUDE HOLD MODE

This longitudinal autopilot mode produces a pitch command input to maintain a commanded altitude. The pitch command is generated by:

$$\delta_{col} = 0.0007\left[\frac{(H - HC)}{30.0} + 5.0\,\dot{z}\right] + DCZ$$

Where: H = present altitude (meters, MSL)

HC = commanded altitude (meters, MSL)

$\dot{z}$ = vertical velocity (m/sec)

DCZ = neutral stick position for pitch

This function compares present altitude to the desired altitude, adjusts for the rate of climb or descent, and then adds its signal to the "neutral stick" signal. In this mode, any pilot pitch input is ignored. However, the pilot must still control airspeed. If the pilot allows the airspeed to get low and has a low power setting, the aircraft will stall.

A-2.2.5 SPEED HOLD MODE

This autopilot mode tries to maintain the airspeed present at mode engagement. This mode does not affect stick inputs and the pilot has complete longitudinal and lateral control. However, the pilot has no throttle control. The system will attempt to maintain airspeed within the limits of idle throttle to full throttle. The Laplacian autopilot command equation is:

$$\delta_{ct} = -0.33\left[\frac{v}{1.0 + 2.0s}\right] - VC$$

Where: v = airspeed (m/sec)

VC = commanded airspeed (m/sec)

This generates a throttle adjustment command which is smoothed by a first order lag. The lag simulates turbine engine response.

A-2.2.6 SPEED AND ALTITUDE HOLD MODE

This mode combines the previously described speed hold and altitude hold modes. Pilot pitch and throttle inputs are ignored, but lateral inputs are unaffected.

A-2.2.7 GLIDE SLOPE COUPLE PLUS SPEED HOLD MODE

When the pilot selects this mode, speed hold is engaged for the aircraft's airspeed at the engagement time. In addition, the autopilot tries to capture the glide slope if an ILS is selected.

If the aircraft is further than 5.4 nm (10 km) from the outer marker, the autopilot goes into a speed and altitude hold mode until within range. Once within range, the system calculates a reference glidepath. This reference glidepath provides a correction to the actual glideslope proportional to the distance to the runway.

If the aircraft height is more than 30.0 m (98.4 ft) AGL, the reference glidepath equation is:

$$GREF = G\emptyset - 0.0012\ D*GSE$$

Where: $G\emptyset$ = simulator glide slope angle: -3.0°, -0.05236 radians

D = distance to the runway

GSE = Glide Slope error

The preceding equation acts within a limit. GREF is calculated to provide a glidepath which corrects to the desired 3° glideslope as a function of error and distance from the runway. However, if the product of the glide slope error and distance from the runway exceeds a certain value and the aircraft is below the glideslope, the autopilot goes into an altitude hold mode.

If the aircraft is above 30.0 meters AGL and within glideslope error limits, the pitch command correction signal is:

$$\delta_{col} = -0.8 \left[GREF - \frac{\dot{z}}{V} \right]$$

The pitch command is proportional to the difference between the reference glide path and the current horizontal velocity.

If the aircraft is below 30.0 meters AGL, the autopilot attempts to land the aircraft and a different GREF relationship is used:

$$GREF = -0.0015 \ ALT + 0.005$$

Where: ALT = height AGL (meters)

This equation produces a "flaring" glidepath to land the aircraft. This GREF is used by the preceeding equation to obtain δ_{col}.

REFERENCES

1. Bailey, R. W.. Human performance engineering: a guide for system designers. New Jersey: Prentice-Hall, Inc.; 1982.

2. Casali, J. G.; Wierwille, W. W.. Communications-imposed pilot workload: a comparison of sixteen estimation techniques. Paper presented at the second symposium on aviation psychology; 1983 April 25-27; Columbus, Ohio.

3. Chu, Y. Y.; Rouse, W. B.. Adaptive allocation of decision making responsibility between human and computer in multi-task situations. IEEE Transactions on Systems, Man, and Cybernetics. SMC-9 (No. 12): 769-778; 1979.

4. Crow, E. L.; Davis, F. A.; Maxfield, M. W.. Statistics manual. New York: Dover Publications, Inc.; 1960.

5. Hart, S. G.; Bortolussi, M. R.. Pilot errors as a source of workload. Paper presented at the second symposium on aviation psychology; 1983 April 25-27; Columbus, Ohio.

6. Heffley, R.K.; Jewell, W. F.. Aircraft handling qualities data. NASA CR-2144. 1972 December.

7. Hicks, J.A.; Soliday, S. M.. An evaluation of sinus arrythmia as a measure of mental load. Knowles, W. B.; Sanders, M; Muckler, F., eds. Proceedings of the sixteenth annual meeting of the Human Factors Society; 1972: 191-196.

8. Hicks, T.G.; Wierwille, W. W.. Comparison of five mental workload assessment procedures in a moving-base driving simulator. Human Factors. 21(2): 129-143; 1979.

9. Jensen, R. S.; Chappell, S.. Pilot performance and workload assessment: an analysis of pilot errors. Report submitted to NASA Ames Research Center in 1983 February.

10. Johannsen. G.. Workload and Workload Measurement. N. Moray, ed. Mental workload; theory and measurement. New York: Plenum; 1979.

11. Kantowitz, B. H.; Hart, S. G.; Bortolussi, M. R.. Measuring pilot workload in a motion-based simulator: asynchronous secondary choice-reaction task. Paper submitted to the IEEE Transactions on Systems, Man, and Cybernetics. 1983.

12. Katz, J. G.. Pilot workload in the air transport environment: measurement, theory, and the influence of Air Traffic Control. Massachusetts Institute of Technology Flight Transportation Laboratory. FTL Report R80-3. 1980 May.

13. Kleinman, D. L.; Baron, S.; Levison, W. H.. A control theoretic approach to manned-vehicle systems analysis. IEEE Transactions on Automatic Control. AC-16: 824-832; 1971.

14. Leplat, J.. Factors determining workload. Ergonomics. 21: 143-149; 1978.

15. Loftus, G. R.; Dark, V. J.; Williams, D.. Short-term memory factors in ground controller/pilot communication. Human Factors. 21(2): 169-181; 1979.

16. McRuer, D.; Ashkenas, I.; Graham, D.. Aircraft dynamics and automatic control. New Jersey: Princeton University Press; 1973.

17. Miller, G. A.. The magical number seven, plus or minus two: some limits on our capacity for processing information. Psychological Review. 63: 81-97; 1956.

18. Rasmussen, J.. Outlines of a hybrid model of the process plant operator. Sheridan, T. B.; Johannsen, G., eds. Monitoring behavior and supervisory control. New York: Plenum; 1976.

19. Rehmann, J. T.. Cockpit display of traffic information and the measurement of pilot workload: an annotated bibliography. Atlantic City, New Jersey: FAA Technical Center, Report No. FAA-EM-81-9. 1982 February.

20. Rehmann, J. T.; Stein, E. S.; Rosenberg, B. L.. Subjective pilot workload assessment. Human Factors. 25(3): 297-307; 1983.

21. Rouse, W. B.. Human-computer interaction in multi-task situations. IEEE Transactions on Systems, Man, and Cybernetics. SMC-7 (NO. 5): 384-392; 1977.

22. Sheridan, T. B.; Ferrell, W. R.. Man-machine systems: information, control, and decision models of human performance. Massachusetts: MIT Press; 1974.

23. Sheridan, T. B.; Rouse, W. B.. Supervisory sampling and control: sources of suboptimality in a tracking task. Proceedings of the seventh annual conference on manual control. NASA SP-281. 1971.

24. Sheridan, T. B.; Simpson, R. W.. Toward the definition and measurement of the mental workload of transport pilots. Massachusetts Institute of Technology Flight Transportation and Man-Machine Laboratories, Technical Report No. DOT-OS-70055. 1979 January.

25. Tanaka, K.; Buharali, A.; Sheridan, T. B.. Mental workload in supervisory control of automated aircraft. Proceedings of the nineteenth annual conference on manual control; 1983 May 23-25: 40-58.

26. Tong, R. M.. A control engineering review of fuzzy systems. Automatica. 13 (No. 6): 559-569; 1977.

27. Tulga, M. K.; Sheridan, T. B.. Dynamic decisions and workload in multitask supervisory control. IEEE Transactions on Systems, Man, and Cybernetics. SMC-10 (No. 5): 217-232; 1980.

28. Walden, R. S.; Rouse, W. B.. A queueing model of pilot decision making in a multi-task flight management situation. IEEE Transactions on Systems, Man, and Cybernetics. SMC-8 (No. 12): 867-875; 1978.

29. Watson, G.; McGaw, D.. Statistical inquiry. New York: John Wiley and Sons; 1980.

30. Wierwille, W. W.; Gutmann, J. C.. Comparison of primary and secondary task measures as a function of simulated vehicle dynamics and driving conditions. Human Factors. 20: 233-244; 1978.

31. Williges, R. C.; Wierwille, W. W.. Behavioral measures of aircrew mental workload. Human Factors. 21(5): 549-574; 1979.

32. Wood, W. T.; Sheridan, T. B.. The use of machine aids in dynamic multi-task environments: a comparison of an optimal model to human behavior. Cambridge, Massachusetts: Massachusetts Institute of Technology; 1982. Master's Thesis.

1. Report No.	2. Government Accession No.	3. Recipient's Catalog No.
CR 177388		

4. Title and Subtitle	5. Report Date
Effect of Time Span and Task Load on Pilot Mental Workload	January 1986
	6. Performing Organization Code: FL

7. Author(s)	8. Performing Organization Report No.
Berg, S. L. & Sheridan, T. B.	

9. Performing Organization Name and Address	10. Work Unit No.
Massuchetts Institute of Technology Man Machine Systems Laboratory Cambridge, MA	
	11. Contract or Grant No.: NAG 2-227

12. Sponsoring Agency Name and Address	13. Type of Report and Period Covered
National Aeronautics and Space Administration Washington, D. C. 20546	Contractor Report
	14. Sponsoring Agency Code: 505-67-11

15. Supplementary Notes

Point of contact: Technical Monitor, Sandra G. Hart - MS 239-3
Ames Research Center, Moffett Field, CA 94035
(415) 694 6072

16. Abstract

Two sets of simulations are described that were designed to examine how a pilot's mental workload would be affected by continuous manual-control activity versus discrete mental tasks that included the length of time between receiving an assignment and executing it. A fixed-base flight simulator was used that consisted of a control box (joystick, throttle, switches for operating electronic and mechanical systems) and a high resolution CRT. Aircraft dynamics were modeled on a Lockheed Jetstar business jet. The CRT display consisted of a forward "out-the-window" perspective view and a cockpit instrument/indicator presentation.

The first experiment evaluated two types of measures: objective performance indicators and subjective ratings. Pilots flew a high-workload manual control mission and a high-workload mission that emphasized mental activities. Subjective ratings for the two missions were different, but the objective performance measures (altitude deviations) were similar. IN the second set of experiments, workload levels were increased and a second performance measure was taken (e.g., airspeed deviations). Mental workload had no influence on either performance-based workload measure. Subjective ratings discriminated among the scenarios and correlated with performance measures for high-workload flights. The number of mental tasks performed did not influence error rates, although high manual workloads did increase errors.

17. Key Words (Suggested by Author(s))	18. Distribution Statement
Pilot workload, aircraft simulation	Unclassified, Unlimited STAR Category 54

19. Security Classif. (of this report)	20. Security Classif. (of this page)	21. No. of Pages	22. Price*
Unclassified	Unclassified	150	

*For sale by the National Technical Information Service, Springfield, Virginia 22161